EDITED BY SERGEI NIRENBURG, HAROLD SOMERS, AND YORICK WILKS

READINGS IN MACHINE TRANSLATION

Readings in Machine Translation

Some of the authors and chairmen at the International Machine Translation Conference, 1961. Left to right: Dr. Olga Kulagina (U.S.S.R.), Mr. D. W. Davies (NPL), Dr. Mukhin (U.S.S.R.), Mr. J. McDaniel (NPL), Prof. Silvio Ceccato (Italy), Mr. Itiroo Sakai (Japan), Mr. R. See (U.S.A.), Miss Jehane Barton (Italy), Dr. Gilbert W. King (U.S.A.), Miss Amelia Janiotis (U.S.A.), Prof. S. Comet (Sweden), Prof. Anthony G. Oettinger (U.S.A.), Prof. S. Vauquois (France), Miss Margaret Masterman (U.K.), Prof. V. H. Yngve (U.S.A.), Dr. A. H. Uttley (NPL), and Mr. David G. Hays (U.S.A.).

Readings in Machine Translation

Edited by Sergei Nirenburg, Harold Somers, and Yorick Wilks

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Times New Roman on 3B2 by Asco Typesetters, Hong Kong, and printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Readings in machine translation / edited by Sergei Nirenburg, Harold Somers, and Yorick Wilks.
p. cm.
"A Bradford book."
Includes bibliographical references and index.
ISBN 0-262-14074-8 (hc. : alk. paper)
1. Machine translation. I. Nirenburg, Sergei. II. Somers, H. L. III. Wilks, Yorick, 1939–
P308.R43 2003
418′.02′0285—dc21

2001056217

Source notes can be found on page 407.

10 9 8 7 6 5 4 3 2 1

II THEORETICAL AND METHODOLOGICAL ISSUES

This page intentionally left blank

Introduction

Sergei Nirenburg, Harold Somers, and Yorick Wilks

Machine translation (MT) has recently celebrated its 50th birthday. This is a short life span for a science, but in that period remarkable progress has been made, mirroring the advances in the contributing disciplines of computer science and linguistics. The articles which we have collected here represent—we hope—the most important papers from the past 50 years, starting with Warren Weaver's memorandum, which is widely believed to have set the whole enterprise in motion.

We should clarify at this point what this book *is not*. This is not an introductory textbook on MT. A number of such texts already exist: we might mention Hutchins & Somers (1992) and Arnold et al. (1994), though neither of these texts covers the most recent developments in much depth. Nor is it a review of the history of MT, for which Hutchins' (1986) book has still to be surpassed, though again, as its publication date would suggest, it says nothing about the most recent developments. Overviews which cover recent history can be found in the "state-of-the-art" presentations that are a feature of the biennial MT Summit series, hosted by the International Association for Machine Translation; these conferences usually include a paper devoted to predictions about the near future, and it is interesting to compare these predictions over the years with the reality as reflected in the other presentations. Other sources of information on the latest developments can be found in the proceedings of conference series such as TMI (Theoretical and Methodological Issues in MT) and the recently launched AMTA (Association for Machine Translation in the Americas)—a list of dates and venues of these conferences is given below—and to a lesser extent in Aslib's Translating and the Computer series, held in London every November since 1979. Other conference series at which MT is usually well represented include Coling (sponsored biennially by ICCL, the International Committee on Computational Linguistics) and ACL (Association for Computational Linguistics), founded as the Association for Machine Translation and Computational Linguistics at the 1962 MT conference at Princeton. Another obvious source of information is the field's premier journal *Machine Translation*, published by Kluwer, while other journals such as *Computational Linguistics and Natural Language Engineering* occasionally include MT-related articles. We have not attempted to provide an extensive bibliography. We list below a number of major books on MT, including the textbooks already mentioned, and a number of collections of articles.

Let us turn to the question of what this book *is*. As researchers and teachers in the field of MT, we have been struck by the number of important and much-cited articles, especially older papers, which are *difficult to find*. So we had the simple idea of putting together in one volume the "classical" MT papers that researchers and students want, or should be persuaded, to read. We assume that readers come to this collection with some knowledge about MT and its history: here they will find the articles which make up MT's communal inheritance.

Having decided to put together this collection, the next task was to decide what should be included. Some decisions were easy: historical papers like Warren Weaver's memorandum, Bar-Hillel's "box in the pen" article, something from the ALPAC

report, the first appearance of the "MT pyramid" diagram, and so on, had to be included. The especially active 1980s are well represented starting with Martin Kay's influential "Proper Place" paper, something on the sublanguage approach, the transfer vs. interlingua debate, knowledge-based MT, and in recent times, example-based MT and speech translation.

The collection is divided into three sections. The first contains "historical" papers from MT's early history, up to the late 1960s, and the ALPAC report, which is often taken as a watershed in MT development (though, as our collection shows, many of the ideas of the "2nd generation" of MT system design were around as much as 10 years before the ALPAC report). The second section contains papers addressing theoretical and methodological issues: sublanguage and controlled input, the role of humans in machine-aided translation, the impact of certain linguistic approaches, the transfer vs. interlingua question, the representation of meaning and knowledge. The third section concentrates on system design, and as such overlaps slightly with the previous two sections. Here we find proposals for multilevel analysis and representation, the knowledge-based, statistical and example-based approaches, computational issues, and so on.

Early on in the planning of this collection we made a decision not to include descriptions of individual systems: it would have been difficult to decide which ones to include and which to leave out, and in any case, several still available collections of articles on MT consist essentially of system descriptions either by the system designers themselves (notably King 1987, Slocum 1988) or third-party commentaries (Hutchins & Somers 1992, Whitelock & Kilby 1996). More recently, some systems are covered by a single publication (Copeland et al. 1991a,b, Goodman & Nirenburg 1991, Kay et al. 1994, Rosetta 1994).

Historical significance was the main criterion for inclusion in the volume. For this reason, the articles chosen do not date much beyond the early 1990s: for how could we be sure that something appearing in the last year or two would become significant? A second important criterion was availability: many of the papers included here are either very old, and so difficult to find, or else appeared in obscure collections, or as Technical Reports; yet still they are cited. The third criterion was the personal taste of the editors. Not everyone sees the history of MT in the same way, and certainly the reader will question some of the papers we have included—and some we have omitted. We hope that these will be too few to impede the reader's enjoyment and appreciation of the collection, but if there are any glaring omissions, we would be pleased to hear your opinions. A final criterion—indeed a limiting criterion which we regret to a certain extent, and which is ironic in view of the subject matter of this collection—is that all the papers included here originally appeared in English. Certainly, there are also significant historical papers in Russian from the early era, in Japanese and French from the 1980s, and perhaps in other languages that we have no access to.

Another project (for another editorial team perhaps) would be to make a collection of translations of such papers as David Hays did in the early days[1] for Russian MT. And perhaps we are not so far from the day when such a collection of translations could be produced with the aid of an MT system, or at least an MAT system ... which brings us round full circle to where we started.

The collection has taken us more time than we expected to put together and bring to publication. One problem was tracing the copyright permissions on some of the older papers, and getting publishers' and authors' permissions to include them. We have done our best in this endeavor, as the list of acknowledgments shows. However, we have been unable to identify the copyright owner, or to get a reply from them, in

some cases: if such people would contact us we would be happy to rectify that situation in any future editions.

As you can see, the papers have all been reset, and we have taken this opportunity to edit some of them lightly, correcting spelling mistakes and other minor inaccuracies. In some cases we have abridged the articles a little, though never, we trust, thereby perverting the author's intended message. In any case we have always indicated where a passage has been cut, and, where necessary, included a brief resumé of the abridged portion. We have had to redraw some of the figures and diagrams, and we have added glosses and translations of examples where necessary. We have also harmonized and verified all the bibliographic references. Despite our best efforts, no doubt some errors remain, for which we apologize; we would be pleased to have these pointed out to us for eventual correction.

Putting this collection together has been a labor of love, but a huge labor nonetheless, and one which we could not have completed alone. We would like to thank many people who have helped us along the way, but in particular the following have been especially helpful: Deborah Field, Linda Fresques, Eva Hajičová, John Hutchins, Margaret Jones, Martin Kay, Frank Knowles, Makoto Nagao, Daniel Ponsford, Jennifer Potter, Charlene Shepard, and Akira Shimazu.

Dates and Locations of MT Conferences

Early MT Conferences (Source: Hutchins 1986)

1. MIT, Cambridge, Mass., 17–20 June 1952.
2. Demonstration of Georgetown–IBM system, New York, 7 January 1954.
3. King's College, Cambridge, England, August 1955.
4. International Conference, MIT, Cambridge, Mass., October 1956.
5. Session at UNESCO conference, Paris, France, 15–20 June 1959.
6. National Symposium, UCLA, Los Angeles, February 1960.
7. Wayne State University, Princeton, N.J., July 1960.
8. Second "Princeton-type meeting," Georgetown University, Washington D.C., 1961.
9. National Physical Laboratory, Teddington, England, November 1961.
10. Third "Princeton-type meeting," Princeton, N.J., 1962.
11. NATO Advanced Summer Institute, Venice, Italy, June 1962.
12. Fourth "Princeton-type meeting," Las Vegas, Nev., 1965.

COLING

1. New York, United States, 1965.
2. Grenoble, France, 1967.
3. Stockholm, Sweden, 1969.
4. Debrecen, Hungary, 1971.
5. Pisa, Italy, 1974.
6. Ottawa, Canada, 1976.
7. Bergen, Norway, 1978.
8. Tokyo, Japan, 1980.

9. Prague, Czechoslovakia, 1982.
10. Stanford, Calif., United States, 1984.
11. Bonn, West Germany, 1986.
12. Budapest, Hungary, 1988.
13. Helsinki, Finland, 1990.
14. Nantes, France, 1992.
15. Kyoto, Japan, 1994.
16. Copenhagen, Denmark, 1996.
17. Montreal, Canada, 1998.
18. Luxembourg, 2000.

TMI (Theoretical and Methodological Issues in the Machine Translation of Natural Languages)

1. Colgate University, Hamilton, N.Y., United States, 14–16 August 1985.
2. Carnegie Mellon University, Pittsburgh, Pa., United States, 12–14 June 1988.
3. University of Texas at Austin, Austin, Tex., United States, 11–13 June 1990.
4. CCRIT–CWARC, Laval (Quebec), Canada, 25–27 June 1992.
5. Kyoto International Community House, Kyoto, Japan, 14–16 July 1993.
6. Katholieke Universiteit Leuven, Belgium, 5–7 July 1995.
7. St. John's College, Santa Fe, N. Mex., United States, 23–25 July 1997.
8. University College, Chester, England, 23–25 August 1999.

MT Summit

1. Hakone, Japan, 17–19 September 1987.
2. Munich, West Germany, 16–18 August 1989.
3. Washington, D.C., United States, 1–4 July 1991.
4. Kobe, Japan, 20–22 July 1993.
5. Luxembourg, 10–13 July 1995.
6. San Diego, Calif., United States, 29 October–1 November 1997.
7. Singapore, 13–17 September 1999.

AMTA (Association for Machine Translation in the Americas)

1. Columbia, Md., United States, 5–8 October 1994.
2. Montreal, Quebec, Canada, 2–5 October 1996.
3. Langhorne, Pa., United States, 28–31 October 1998.
4. Cuernavaca, Mexico, 10–14 October 2000.

Note

1. O. S. Akhmanova, I. A. Mel'čuk, R. M. Frumkina, and E. V. Paducheva (1963), Exact Methods in Linguistic Research (trans. D. G. Hays and D. V. Mohr). Rand Corporation Memorandum R-397-PR, Santa Monica, CA.

References

Arnold, D. 1994. *Machine Translation: An Introductory Guide*. Manchester: NCC Blackwell.

Copeland, C., et al. (eds.) 1991a. *The Eurotra Linguistic Specifications*. Luxembourg: Office for Official Publications of the Commission of the European Community.

Copeland, C., et al. (eds.) 1991a. *The Eurotra Formal Specifications*. Luxembourg: Commission of the European Communities, Directorate-General Telecommunications, Information Industries and Innovation.

Goodman, K., and S. Nirenburg (eds.). 1992. *The KBMT Project: A Case Study in Knowledge-Based Machine Translation*. San Mateo, Calif.: Morgan Kaufmann.

Hutchins, W. J. 1986. *Machine Translation: Past, Present, Future*. Chichester: Ellis Horwood.

Hutchins, W. J., and H. L. Somers. 1992. *An Introduction to Machine Translation*. London: Academic Press.

Kay, M., M. Gawron, and P. Norvig. 1994. *Verbmobil: A Translation System for Face-to-Face Dialog*. Stanford, Calif.: Center for the Study of Languages and Information.

King, Margaret (ed.). 1987. *Machine Translation Today: The State of the Art: Proceedings of the Third Lugano Tutorial, Lugano, Switzerland, April 2–7, 1984*. Edinburgh, Scotland: Edinburgh University Press.

Rosetta, M. T. 1994. *Compositional Translation*. Dordrecht, the Netherlands: Kluwer Academic Publishers.

Slocum, J. 1988. *Machine Translation Systems*. Cambridge: Cambridge University Press.

Whitelock, P., and K. Kilby. 1995. *Linguistic and Computational Techniques in Machine Translation System Design*. London: UCL Press.

This page intentionally left blank

I
HISTORICAL

This page intentionally left blank

INTRODUCTION

Sergei Nirenburg

As a research and development field, machine translation (MT) is among the oldest among the various subdisciplines and applications of computer science to the study of natural language. MT is also a subdiscipline of computational linguistics or, one could say, one of the latter's flagship application areas. MT, in fact, historically predates CL and has helped to usher in that field of inquiry and in many ways shaped its early directions and concerns. Indeed, to give just a few examples, the journal *Computational Linguistics*, so familiar to us today, started its existence as *Mechanical Translation*, and was later renamed, in turn, *Mechanical Translation* and *Computational Linguistics* and *The American Journal of Computational Linguistics* before assuming its current name. Also, Prolog, a major programming language, was launched with MT in mind.

While MT is an application area, it is surprising that it can hardly be considered a direct application of theoretical or descriptive linguistics. (A few MT efforts over the years—for instance, Rosetta or Unitran—claimed a theoretical lineage. However, invariably, the theoretical work on which these MT efforts were based had to be modified very seriously, often to the point of evoking the well-known "stone soup" metaphor.) This was painfully obvious in the early days of MT. As was correctly noted by Erwin Reifler as early as 1955 in the article reproduced in this collection,

> The MT linguist [...] will be mostly concerned with differences in behavior between a given pair of languages. He need not adhere strictly to the results of scientific language research. When they serve his purpose, he will consider them. But he will ignore them when an arbitrary treatment of the language material better serves his purpose [...] Practicality, for the MT linguist, is a consideration of the highest order. [...] MT is concerned primarily with meaning, an aspect of language that has often been treated as a poor relation by linguists and referred to psychologists and philosophers.

From the 1960s on, MT was, in fact, often used to apply contemporary linguistic theories, but the systems that were directly inspired by a particular linguistic theory were usually seldom comprehensive or broad-coverage. The discrepancy between the needs of MT and the goals and theories in linguistics is real, and the relationships between MT and the would-be primary "natural" source of inspiration for MT research are still not very close. As to other influences, MT has been an eclectic area where a variety of methods were attempted, from language descriptions based on "first principles" to influences from knowledge representation within the field of artificial intelligence (another area which MT arguably helped launch) to stochastic methods imported from information theory and mathematical statistics to artificial neural nets. Parallel to the search for the best underlying method for carrying out translation was the policy to use the best and newest advances in computer hardware and software.

Before the advent of the digital computer, building a machine to translate among human languages was more or less in the realm of science fiction, though this did not stop the Soviet engineer Petr Smirnov-Trojanskij from patenting, in 1933, a

mechanical device for, essentially, storing and using multilingual dictionaries or from continuing for more than 15 years to work on mechanical translation on the basis of this device. MT has been widely considered a tangible goal since the late 1940s, with the advent of the digital computer, the concept of stored program and the promise of large storage devices. Translation among languages[1] was among the first non-numerical applications suggested and actually attempted for the nascent computing technology.

Why exactly MT has become such a high-profile area so early is not clear. Certainly, the wartime successes of cryptography in the early 1940s in the U.K. and U.S. had an influence on this. The mathematicians and early computer scientists who made spectacular progress in breaking the enemy codes during the war undertook, riding the wave of spectacular successes, to branch into other endeavors and extend their methods, proven on a complex task, to other areas. Importing a technique or a theory that proved successful or promising in one area into another has always been popular. Thus, in the second half of the 19th century the German Young Grammarians investigated historical rules of development of languages under the influence of Darwin's theory and in the 1920s Sapir and Whorf worked on the "theory of linguistic relativity." Similarly, in the past decade statistical methods were used in research on the human genome, in machine translation and in predicting stock market behavior.

Translation of natural language seemed to be a very natural extension for the methods used in breaking codes. It is no surprise, therefore, that the treatise universally considered as the major impetus for the original interest in MT proceeds intellectually from the metaphor of cryptography: in his famous memorandum, Warren Weaver states: "One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: 'This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.'"

As will be made clear from the texts of the contributions in this section, the MT pioneers were always aware of the applied nature of their field. The ideas and techniques imported from other fields were all to serve the immediate and practical goals of building MT systems.[2] In the 1960s, the field gradually became much more method-oriented, and many (though definitely not all) projects, while paying lip service to the practical needs of MT, would concentrate much more on applying and testing a variety of linguistic (e.g., syntactic) and computational linguistic (e.g., parsing) theories within the framework of MT. The pendulum would swing once again in the late 1980s, when the renewed emphasis on results and system evaluation in competition would bring back the engineering methods and attitudes familiar from the early days of MT and often quite detached from the knowledge accumulated in linguistics.

It is, indeed, remarkable how little impact theoretical linguistics had on the early machine translation. The new discipline borrowed more not only from cryptography but also from philosophy and mathematical logic. Indeed, Yehoshua Bar-Hillel, widely credited for being the first person appointed to work in MT proper (at the MIT Research Laboratory for Electronics, in 1951) was a mathematical logician and a philosopher. In fact, there is a lot of weight to the claim that the work on MT led to the birth of computational linguistics and artificial intelligence (or, at least, its natural language processing component).

Knowledge of the history of one's area of endeavor is indispensable for a scholar, even in technological fields, where often a system or a device is rendered antiquated by new research and development efforts very soon after it is implemented and deployed. As MT is not even a purely technological field, that awareness of approaches,

opinions and methods past can and should be of direct practical help to workers in the field.

Fortunately, machine translation has found a dedicated and prolific "chief archivist" in John Hutchins. In his book *Machine Translation: Past*, *Present*, *Future* (1986) and in the 58-page historical survey article in *Machine Translation* in 1997, he presents a vivid general picture of the events surrounding the early developments in machine translation. His comments on the ALPAC report, a crucial juncture in the history of MT research, are reproduced in this collection (part I, "ALPAC: The (In)famous Report").

The contributions collected in the historical part of this collection are intended to give the reader an idea about how vibrant the research in MT was in its early days (roughly, from its inception in the late 1940s till 1965); how many of the still current approaches and methods were first proposed and tried in those times; and how diligently many of the contributors and their groups worked on practical implementations of their ideas within the confines of their contemporary technology (with no high-level computer languages; no interactive terminals, to say nothing of graphical user interfaces; no online resources; with machines whose memories were smaller than those of contemporary hand-held calculators, etc.).

Over the course of the roughly 15 years covered by the contributions in this part, technology made great strides forward, for indeed it is difficult to see how a complex and large-scale MT program, such as, for instance, described in the contribution by Ida Rhodes (part I, "A New Approach to Mechanical Syntactic Analysis of Russian"), could be developed using the techniques reported by Andrew Booth in his contribution describing the earliest experiments in MT more than a decade earlier (part I, "Mechanical Translation").

The contributions in this section are in approximate chronological order. Just as in Locke and Booth (1955), the first collection of MT articles published in book form, Warren Weaver's memorandum opens our reader (part I, "Translation"). The story of the memorandum and the events that both led to it and followed it is well presented in Hutchins (1997). If ever there was a case of a well-informed, well-positioned and forward-looking enthusiast almost single-handedly creating the initial momentum for a discipline, it is Warren Weaver with respect to MT. He energized the early MT research, not least through his influence on the funding priorities at the National Science Foundation of the United States. Thus, among other recipients of early grants to carry out experiments in non-numerical applications of computing was Andrew Booth of Birkbeck College of the University of London, who concluded, in late 1947, that MT was a prime area for such an endeavor. The contribution by Booth in this collection (part I, "Mechanical Translation") describes some of his early experimental settings and ideas about MT. It is a very interesting document in that the reader should realize that the work described was truly trail-blazing and pioneering. There was no paradigm of MT research in existence yet, and even though Booth does not present his work in a paradigmatic mode, some tacit assumptions about it are interesting to note.

While Booth's approach is strictly practical and based on first principles, the contribution by Erwin Reifler (part I, "The Mechanical Determination of Meaning"), an influential early MT researcher, casts a wider methodological net and tries to suggest some generalizations and abstractions about the process of translation, as well as some connections with and differences from research in linguistics.

Thus, the following observation about the process of translation sets up the overall view of MT as a process of ambiguity resolution. "A complete message contains information that, together with a certain number of unsymbolized situational criteria, enables the human hearer, reader, or translator to select the intended meanings

from the multiple potential meanings characterizing its constituents.'' Reifler quotes Bloomfield: ''... as to denotation, whatever can be said in one language can doubtless be said in any other ... the difference will concern only the structure of the forms, and their connotation'' to stress that the basis of translation is in the invariance of meaning across languages. Already in the early 1950s it was clear to Reifler that high-quality translation must take into account metaphors, metonymies, similes and other non-literal language phenomena:

> The determination of intended meaning depends not only on the semantic peculiarities of the source language, but on the semantic peculiarities of the target language as well! As already mentioned, our problem is multiple meaning in the light of source-target semantics. If, for instance, we want to translate the English sentence, ''He is an ass,'' into Chinese, we must discover whether the Chinese word for ''ass'' can be used as a contemptuous expression denoting a stupid human being. As a matter of fact, it cannot be so used, and therefore a literal translation would be completely unintelligible. Another Chinese word meaning something like ''stupid'' or ''foolish'' has to be substituted or else the English sentence has to be expressed in a completely different way according to the idiomatics of the Chinese language.

Of course, most of the present-day MT systems do not attempt to resolve this type of problem dynamically, and typically are only capable of doing this (or even considering this as a problem!) if the appropriate reading is listed among the senses in the transfer dictionary.

Another interesting find is the following early statement concerning, essentially, the issue of selectional restrictions:

> From among multiple nongrammatical meanings the translation mechanism will extract the intended meaning by determining the nongrammatical meaning in which two or more syntactically correlated source forms coincide. For example, in *Er bestand die Prüfung* (he passed the examination) the memory equivalent of *bestand* will be accompanied by a number of distinctive code signals, each indicative of one of its multiple nongrammatical meanings. One of these code signals will be identical with a code signal accompanying the memory equivalents of all substantives which, as objects of *bestand*, ''pinpoint'' the intended meaning of the latter as one best translated by English ''passed.''

The stochastic approach to MT had its beginnings not in the late 1980s, as many believe, but thirty years earlier. The short contribution by Gil King (part I, ''Stochastic Methods of Mechanical Translation'') is ample evidence of that. King envisaged an environment in which stochastic techniques were used for disambiguating among the candidate translations of source language words, while the rest of the system was built using ''traditional'' dictionaries and processors. Here are some statements that set forth the motivation of King's approach:

> It is well known that Western languages are 50% redundant. Experiment shows that if an average person guesses the successive words in a completely unknown sentence he has to be told only half of them ... a machine translator has a much easier problem—it does not have to make a choice from the wide field of all possible words, but is given in fact the word in the foreign language, and only has to select one from a few possible meanings.
>
> In machine translation the procedure has to be generalized from guessing merely the *next* word. The machine may start anywhere in the sentence and skip around looking for clues. The procedures for estimating the probabilities and selecting the highest may be classified into several types, depending on the type of hardware in the particular machine-translating system to be used.

The contribution by Victor Yngve (part I, ''A Framework for Syntactic Translation'') belongs to the wave of MT efforts that followed the initial experimentation. It represents more mature research activities that led the field to deeper and more

comprehensive descriptions of the requirements and approaches to MT. Yngve's paper enumerates types of clues for source text analysis, anticipating the central issues of the area of natural language parsing. It also introduces an influential discussion of the "100%" vs. "95%" approaches to MT:

The six types of [analysis] clues are

1. The field of discourse.
2. Recognition of coherent word groups, such as idioms and compound nouns.
3. The syntactic function of each word.
4. The selectional relations between words in open classes, that is, nouns, verbs, adjectives, and adverbs.
5. Antecedents. The ability of the translating program to determine antecedents will not only make possible the correct translation of pronouns, but will also materially assist in the translation of nouns and other words that refer to things previously mentioned.
6. All other contextual clues, especially those concerned with an exact knowledge of the subject under discussion. These will undoubtedly remain the last to be mechanized. Finding out how to use these clues to provide correct and accurate translations by machine presents perhaps the most formidable task that language scholars have ever faced.

Attempts to learn how to utilize the above-mentioned clues have followed two separate approaches. One will be called the "95 percent approach" because it attempts to find a number of relatively simple rules of thumb, each of which will translate a word or class of words correctly about 95 percent of the time, even though these rules are not based on a complete understanding of the problem. This approach is used by those who are seeking a short-cut to useful, if not completely adequate, translations. The other approach concentrates on trying to obtain a complete understanding of each portion of the problem so that completely adequate routines can be developed.

The name of Yehoshua Bar-Hillel (part I, "The Present Status of Automatic Translation of Languages") is arguably the most famous among all researchers in MT. In view of this, it is remarkable that Bar Hillel, an eminent philosopher of language and mathematical logician, has never written or designed an MT system. In MT, he was a facilitator and an outstanding intellectual critic. His unusual ability to understand the nature of the various problems in MT and the honesty and even-handedness of his—usually very strongly held—opinions set him apart from the run-of-the-mill system designer, too busy building a system to be able fully to evaluate its worth, or amateur critic who often judges MT by an impossible, though popular standard of the best translations performed by teams of professional human translators, editors, domain specialists and proofreaders. The following sample of Bar Hillel's opinions (taken from his article in this reader) will demonstrate how uncannily modern many of them sound.

On the 95 percent approach:

It is probably proper to warn against a certain tendency which has been quite conspicuous in the approach of many MT groups. These groups, realizing that FAHQT [Fully automated, high-quality MT] is not really attainable in the near future so that a less ambitious aim is definitely indicated, had a tendency to compromise in the wrong direction for reasons which, though understandable, must nevertheless be combated and rejected. Their reasoning was something like the following: since we cannot have 100% automatic high-quality translation, let us be satisfied with a machine output which is complete and unique, i.e., a smooth text of the kind you will get from a human translator (though perhaps not quite as polished and idiomatic), but which has a less than 100% chance of being correct. I shall use the expression "95%" for this purpose since it has become a kind of slogan in the trade, with the understanding that it should by no means be taken literally. Such an approach would be implemented by one of the two following procedures: the one procedure would require to print the

most frequent target-language counterpart of a given source-language word whose ambiguity has not been resolved by the application of the syntactical and semantical routines, necessitating, among other things, large scale statistical studies of the frequency of usage of the various target renderings of many, if not most, source-language words; the other would be ready to work with syntactical and semantical rules of analysis with a degree of validity of no more than 95%, so long as this degree is sufficient to insure uniqueness and smoothness of the translation.

On statistics and MT:

No justification has been given for the implicit belief of the "empiricists" that a grammar satisfactory for MT purposes will be compiled any quicker or more reliably by starting from scratch and "deriving" the rules of grammar from an analysis of a large corpus than by starting from some authoritative grammar and changing it, if necessary, in accordance with analysis of actual texts. The same holds *mutatis mutandis* with regard to the compilation of dictionaries.

On context and ambiguity resolution:

It is an old prejudice, but nevertheless a prejudice, that taking into consideration a sufficiently large linguistic environment as such will suffice to reduce the semantical ambiguity of a given word. Why is it that a machine with a memory capacity sufficient to deal with a whole paragraph at a time, and a syntactico-semantic program that goes, if necessary, beyond the boundaries of single sentences up to a whole paragraph (and, for the sake of the argument, up to a whole book)—something which has so far not gotten beyond the barest and vaguest outlines—is still powerless to determine the meaning of *pen* in our sample sentence within the given paragraph?

[Here Bar Hillel refers to his famous example of the text "*Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.*" where the word "pen" cannot be disambiguated between the writing implement and enclosure senses without the use of extralinguistic knowledge about the typical relative sizes of boxes and pens (in both senses).—Eds.]

The contribution by Ida Rhodes (part I, "A New Approach to the Mechanical Syntactic Analysis of Russian") is a very well reasoned and meticulously argued presentation of results of practical MT system development, with a realistic perspective on the complexities of the task at hand. First of all, Rhodes forcefully describes the objective obstacles in the path of a translator, even a human translator, let alone a computer program. She elegantly concludes that

It would seem that characterizing a sample of the translator's art as a good translation is akin to characterizing a case of mayhem as a good crime: in both instances the adjective is incongruous. If, as a crowning handicap, we are asked to replace the vast capacity of the human brain by the paltry contents of an electronic contraption, the absurdity of aiming at anything higher than a crude practical translation becomes eminently patent.

The above makes it clear that "[t]he heartbreaking problem which we face in mechanical translation is how to use the machine's considerable speed to overcome its lack of human cognizance." Rhodes then proceeds to describe the needs of automatic syntactic analysis. It is remarkable how "modern" is her evaluation of the differences between published dictionaries and lexicons (she calls them glossaries) for MT. She then proceeds to describe, in detail, a complex procedure for syntactic analysis of Russian.

The contribution by Susumu Kuno (part I, "A Preliminary Approach to Japanese–English Automatic Translation") describes a method for Japanese–English MT, with an original Japanese segmentor and syntactic analysis following the method of Rhodes. At the time of publication, the method was not yet implemented in a computer system, but it describes the first attempt at solving a very

important problem in processing Asian languages (and other languages with no breaks between words) that has achieved some prominence in the late 1980s and in the 1990s.

The contribution by Sydney Lamb (part I, ''On the Mechanization of Syntactic Analysis'') seems to be a prolegomenon to the currently very fashionable studies devoted to inducing syntactic grammars from corpora and will give a historical perspective for this type of activity.

The contribution by David Hays (part I, ''Research Procedures in Machine Translation''), a leader in the field of MT and computational linguistics in the 1960s and 1970s and one of the founders of the COLING conferences, is mostly interesting for its acute methodological observations concerning the research tasks to be carried out by MT developers. Here is a small sampling:

> Whereas mathematical systems are defined by their axioms, their explicit and standard rules, natural languages are defined by the habits of their speakers, and the so-called rules are at best reports of those habits and at worst pedantry.
>
> Until computational linguistics was conceived, no one needed a fully detailed account of any language for any purpose.
>
> It seems inevitable that text must supersede the informant when the details are to be filled in, simply because no one knows every particular of his language.

We include in this collection excerpts from the 1966 ALPAC report and a commentary on the report and its impact written by John Hutchins (part I, ''ALPAC: The (In)Famous Report''). The report has exerted monumental influence on the development of MT in the U.S.. It is very important for the present-day MT researcher to understand what ALPAC actually said because what usually trickles down the collective memory is only the extra-scientific consequences of its publication, most of all the steep drop in the levels of funding of MT in the US after ALPAC's publication. Reading and discussing this report will clarify certain persistent misconceptions.

The contribution by Silvio Ceccato (part I, ''Correlational Analysis and Mechanical Translation'') is one of the most original ones in this volume. The famous Italian linguist presents a study elegant in style and intriguing in substance; among other reasons, this is because the author does not seem to be influenced, to any significant degree, by the MT scholarship that had been accumulated by the time this contribution appeared. While this might be considered a drawback, it also leads to an original point of view that will help us to present the MT scene as a complex and diverse phenomenon that it was. Here are some of Ceccato's opinions. Echoing Rhodes' position concerning MT glossaries, Ceccato avers that ''the entrepreneurs of mechanical translation must have been unpleasantly surprised for grammar, as it was conceived for men, is not immediately applicable to machines.'' He explains it in an idiosyncratic way, saying that computational grammars are not conceived as links between morphology and semantics.

> The dearth of explicit information, if it does not create difficulties for man, but rather assures him an economic and quick discourse, is troublesome both when he wants to find an algorithm which describes language, and when he wants to mechanize our linguistic activity, and in particular our comprehension of language. We must, in fact, prepare a system of linguistics which distinguishes that which, in the relationship between thought and language, appears explicitly from that which implicitly enters into it.

The above can, in fact, be construed as an argument for an ontology-based approach to language processing!

The contribution by Kulagina and Mel'čuk (part I, "Automatic Translation: Some Theoretical Aspects and the Design of a Translation System") is a bold and surprisingly modern programmatic statement about how one should understand the problem of MT and its "ecology." In their own words:

> Three problems are stated on whose solution, in the writers' view, the successful development of AT [automatic translation] is largely dependent: the linguistic problem (correlation 'text-meaning'), the gnostical problem (correlation 'meaning-reality') and the problem of automating scientific research. . . . For AT needs an algorithmic analogue of this ability to perform the transition from text to its meaning ('T → M') and vice versa ('M → T').

Note that the authors consider meaning extraction a condition *sine qua non* for MT: "three things are required: a means of recording meaning (a special notation), an algorithm of analysis, and of synthesis." The authors do not stress the knowledge requirements for the system.

"Though, historically, the above tasks have first been faced and strictly formulated within AT, they are, in our opinion, tasks of general linguistics, moreover cardinal problems of any serious theory of language." The above is an important statement concerning the goals of theoretical linguistics.

The following is as succinct formulation as any of the dependence of high-quality machine translation on the knowledge of the world:

> Understanding the "linguistic" meaning of a text does not guarantee the ability to process this text correctly: "linguistic" meaning and "situational" content (the state of affairs) are quite different things not always linked by a unique (one-to-one) correspondence. The right translation is possible only if the extralinguistic situation is rightly understood.

And also:

> Any substantial progress of AT is closely dependent on progress in the study of human thinking and cognition, in particular—on the successful solution of such tasks as developing a formal notation for recording external world situations and constructing models of thinking (meaning analysis and synthesis).

Anticipating "naive physics" by at least a decade, accurately down to the term itself, the authors state:

> Of all real situations only very few (highly special, hardly occurring in everyday practice) are described by exact sciences. However, even in scientific texts, not to speak of fiction or journalism, there are many, in no way special, everyday situations whose description and classification seem to be largely (if not absolutely) ignored so far. It is high time that description of such situations became the object of a special branch of science. In other words, we must proceed to build up a regular encyclopedia of the man-in-the-street's knowledge about the everyday world, or a detailed manual of naive, home-spun "physics" written in an appropriate technical language.

Finally, the authors offer an analysis of the types of problems that must be solved for MT to be successful and state that work in MT should continue even while those problems still await an adequate solution. In the rest of the paper, the authors discuss the design of an MT system based on meaning, with an analysis module, a semantic dictionary and a synthesis module. The latter is described in detail, and would be of special interest to researchers in natural language generation. The former are described in rather programmatic terms, but a number of interesting theoretical and methodological points are made. Among other things, the authors talk about translating a source language into its "basic" form and then translating that basic form into a basic form of the target language, off of which the idiomatic form of the text in the target language will be generated.

A similar topic is central to the article selected from the writings of Margaret Masterman (part I, "Mechanical Pidgin Translation"), an MT researcher and teacher of many other luminaries in MT and AI, including Martin Kay and Yorick Wilks:

> There are two lines of research which highlight this problem [...] (1) matching the main content-bearing words and phrases with a semantic thesaurus [...] which determines their meanings in context; (2) word-for-word matching translation into a "pidgin-language" using a very large bilingual word-and-phrase dictionary.

Masterman and her colleagues researched the semantic thesaurus in some detail, and it might be said that that was the original work concerning semantic interlinguas (as opposed to syntactic ones like the one suggested by Vauquois[3]). This work found further development, for instance, in the work of Sparck Jones and Wilks. The paper selected for this collection describes a method of automatically transforming results of low-quality word-for-word MT (with a morphological analyzer!) into a readable form, essentially by carrying out feature transfer between source and target languages. The paper calls for more attention to what the author calls "bits of information" and we would call grammatical morphemes and closed-class lexical elements of a language. The good example of how much these elements contribute to the understanding of the meaning of text is, as Masterman mentions, a text like Lewis Carroll's "Jabberwocky," in which all open-class lexical items are not English, while all the closed class items are.

The paper by Takahashi *et al.* (part I, "English-Japanese Machine Translation") is the first report about the Japanese efforts in MT, which flowered so richly in the 1980s. The paper describes an experiment of translating from English to Japanese some parts of a Japanese textbook of English. A notable feature of this experiment is the use of a specially constructed computer, Yamato. The design of the machine is described, as well as the structure of the 2,000-entry English word dictionary, an English phrasal dictionary (whose size was not mentioned), a syntax "dictionary" which is, in fact, a set of syntactic grammar rules, and the Japanese dictionary.

In preparing the articles for publication in this collection, some parts of these contributions were omitted, partly because they included material which is less instructive to present-day readers or somewhat obsolete and partly simply due to space limitations. The lacunae are marked by [...].

Notes

1. It was only later that the term of choice would become "natural language"—as there were no computer languages of note at the time, and nobody in the sciences paid much attention to artificial languages built for human use, such as Esperanto. Well, nobody at that time would think of calling a guitar an acoustic guitar either.

2. This is, of course, a simplification. Even in the early years of MT there was a division between the "brute-force" and "scientific" approaches. However, the general tenor of the times was undeniably empirical.

3. B. Vauquois, Langages artificiels, systèmes formels et traduction automatique, in A. Ghizetti (ed.), *Automatic Translation of Languages: Papers Presented at NATO Summer School, Venice, July 1962* (Oxford: Pergamon, 1966).

This page intentionally left blank

1
Translation

Warren Weaver

There is no need to do more than mention the obvious fact that a multiplicity of languages impedes cultural interchange between the peoples of the earth, and is a serious deterrent to international understanding. The present memorandum, assuming the validity and importance of this fact, contains some comments and suggestions bearing on the possibility of contributing at least something to the solution of the world-wide translation problem through the use of electronic computers of great capacity, flexibility, and speed.

The suggestions of this memorandum will surely be incomplete and naive, and may well be patently silly to an expert in the field—for the author is certainly not such.

A War Anecdote—Language Invariants

During the war a distinguished mathematician whom we will call *P*, an ex-German who had spent some time at the University of Istanbul and had learned Turkish there, told W. W. the following story.

A mathematical colleague, knowing that *P* had an amateur interest in cryptography, came to *P* one morning, stated that he had worked out a deciphering technique, and asked *P* to cook up some coded message on which he might try his scheme. *P* wrote out in Turkish a message containing about 100 words; simplified it by replacing the Turkish letters ç, ğ, ı, ö, ş, and ü by c, g, i, o, s, and u respectively; and then, using something more complicated than a simple substitution cipher, reduced the message to a column of five-digit numbers. The next day (and the time required is significant) the colleague brought his result back, and remarked that they had apparently not met with success. But the sequence of letters he reported, when properly broken up into words, and when mildly corrected (not enough correction being required really to bother anyone who knew the language well), turned out to be the original message in Turkish. The most important point, at least for present purposes, is that the decoding was done by someone who did not know Turkish, and did not know that the message was in Turkish. One remembers, by contrast, the well-known instance in World War I when it took our cryptographic forces weeks or months to determine that a captured message was coded from Japanese; and then took them a relatively short time to decipher it, once they knew what the language was.

During the war, when the whole field of cryptography was so secret, it did not seem discreet to inquire concerning details of this story; but one could hardly avoid guessing that this process made use of frequencies of letters, letter combinations, intervals between letters and letter combinations, letter patterns, etc., *which are to some significant degree independent of the language used*. This at once leads one to suppose that, in the manifold instances in which man has invented and developed languages, there are certain invariant properties which are, again not precisely but to some statistically useful degree, common to all languages.

This may be, for all I know, a famous theorem of philology. Indeed the well-known *bow-wow*, *woof-woof*, etc. theories of Müller and others, for the origin of languages, would of course lead one to expect common features in all languages, due to their essentially similar mechanism of development. And, in any event, there are obvious reasons which make the supposition a likely one. All languages—at least all the ones under consideration here—were invented and developed by *men*; and all men, whether Bantu or Greek, Icelandic or Peruvian, have essentially the same equipment to bring to bear on this problem. They have vocal organs capable of producing about the same set of sounds (with minor exceptions, such as the glottal click of the African native). Their brains are of the same general order of potential complexity. The elementary demands for language must have emerged in closely similar ways in different places and perhaps at different times. One would expect wide superficial differences; but it seems very reasonable to expect that certain basic, and probably very non-obvious, aspects be common to all the developments.

It is just a little like observing that trees differ very widely in many characteristics, and yet there are basic common characteristics—certain essential qualities of "tree-ness,"—that all trees share, whether they grow in Poland, or Ceylon, or Colombia. Furthermore (and this is the important point), a South American has, in general, no difficulty in recognizing that a Norwegian tree *is* a tree.

The idea of basic common elements in all languages later received support from a remark which the mathematician and logician Reichenbach made to W. W. Reichenbach also spent some time in Istanbul, and, like many of the German scholars who went there, he was perplexed and irritated by the Turkish language. The grammar of that language seemed to him so grotesque that eventually he was stimulated to study its logical structure. This, in turn, led him to become interested in the logical structure of the grammar of several other languages; and, quite unaware of W. W.'s interest in the subject, Reichenbach remarked, "I was amazed to discover that, for (apparently) widely varying languages, the basic logical structures have important common features." Reichenbach said he was publishing this, and would send the material to W. W.; but nothing has ever appeared.

One suspects that there is a great deal of evidence for this general viewpoint—at least bits of evidence appear spontaneously even to one who does not see the relevant literature. For example, a note in *Science*, about the research in comparative semantics of Erwin Reifler of the University of Washington, states that "the Chinese words for 'to shoot' and 'to dismiss' show a remarkable phonological and graphic agreement." This all seems very strange until one thinks of the two meanings of "to fire" in English. Is this only happenstance? How widespread are such correlations?

Translation and Computers

Having had considerable exposure to computer design problems during the war, and being aware of the speed, capacity, and logical flexibility possible in modern electronic computers, it was very natural for W. W. to think, several years ago, of the possibility that such computers be used for translation. On March 4, 1947, after having turned this idea over for a couple of years, W. W. wrote to Professor Norbert Wiener of Massachussetts Institute of Technology as follows:

One thing I wanted to ask you about is this. A most serious problem, for UNESCO and for the constructive and peaceful future of the planet, is the problem of translation, as it unavoidably affects the communication between peoples. Huxley has recently told me that they are appalled by the magnitude and the importance of the translation job.

Recognizing fully, even though necessarily vaguely, the semantic difficulties because of multiple meanings, etc., I have wondered if it were unthinkable to design a computer which would translate. Even if it would translate only scientific material (where the semantic difficulties are very notably less), and even if it did produce an inelegant (but intelligible) result, it would seem to me worth while.

Also knowing nothing official about, but having guessed and inferred considerable about, powerful new mechanized methods in cryptography—methods which I believe succeed even when one does not know what language has been coded—one naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian, I say: "This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."

Have you ever thought about this? As a linguist and expert on computers, do you think it is worth thinking about?

Professor Wiener, in a letter dated April 30, 1947, said in reply:

Second—as to the problem of mechanical translation, I frankly am afraid the boundaries of words in different languages are too vague and the emotional and international connotations are too extensive to make any quasimechanical translation scheme very hopeful. I will admit that basic English seems to indicate that we can go further than we have generally done in the mechanization of speech, but you must remember that in certain respects basic English is the reverse of mechanical and throws upon such words as *get* a burden which is much greater than most words carry in conventional English. At the present time, the mechanization of language, beyond such a stage as the design of photoelectric reading opportunities for the blind, seems very premature. . . .

To this, W. W. replied on May 9, 1947:

I am disappointed but not surprised by your comments on the translation problem. The difficulty you mention concerning Basic seems to me to have a rather easy answer. It is, of course, true that Basic puts multiple use on an action verb such as *get*. But, even so, the two-word combinations such as *get up*, *get over*, *get back*, etc., are, in Basic, not really very numerous. Suppose we take a vocabulary of 2,000 words, and admit for good measure all the two-word combinations as if they were single words. The vocabulary is still only four million: and that is not so formidable a number to a modern computer, is it?

Thus this attempt to interest Wiener, who seemed so ideally equipped to consider the problem, failed to

produce any real result. This must in fact be accepted as exceedingly discouraging, for, if there are any real possibilities, one would expect Wiener to be just the person to develop them.

The idea has, however, been seriously considered elsewhere. The first instance known to W. W., subsequent to his own notion about it, was described in a memorandum dated February 12, 1948, written by Dr. Andrew D. Booth, who, in Professor J. D. Bernal's department in Birkbeck College, University of London, had been active in computer design and construction. Dr. Booth said:

> A concluding example, of possible application of the electronic computer, is that of translating from one language into another. We have considered this problem in some detail, and it transpires that a machine of the type envisaged could perform this function without any modification in its design.

On May 25, 1948, W. W. visited Dr. Booth in his computer laboratory at Welwyn, London, and learned that Dr. Richens, Assistant Director of the Bureau of Plant Breeding and Genetics, and much concerned with the abstracting problem, had been interested with Dr. Booth in the translation problem. They had, at least at that time, not been concerned with the problem of multiple meaning, word order, idiom, etc., but only with the problem of mechanizing a dictionary. Their proposal then was that one first "sense" the letters of a word, and have the machine see whether or not its memory contains precisely the word in question. If so, the machine simply produces the translation (which is the rub; of course "the" translation doesn't exist) of this word. If this exact word is not contained in the memory, then the machine discards the last letter of the word, and tries over. If this fails, it discards another letter, and tries again. After it has found the largest initial combination of letters which *is* in the dictionary, it "looks up" the whole discarded portion in a special "grammatical annex" of the dictionary. Thus confronted by *running*, it might find *run* and then find out what the ending (*n*)*ing* does to *run*.

Thus their interest was, at least at that time, confined to the problem of the mechanization of a dictionary which in a reasonably efficient way would handle *all forms* of all words. W. W. has no more recent news of this affair.

Very recently the newspapers have carried stories of the use of one of the California computers as a translator. The published reports do not indicate much more than a word-into-word sort of translation, and there has been no indication, at least that W. W. has seen, of the proposed manner of handling the problems of multiple meaning, context, word order, etc.

This last-named attempt, or planned attempt, has already drawn forth inevitable scorn, Mr. Max Zeldner, in a letter to the *Herald Tribune* on June 13, 1949, stating that the most you could expect of a machine translation of the 55 Hebrew words which form the 23rd Psalm would start out: *Lord my shepherd no I will lack*, and would close *But good and kindness he will chase me all days of my life; and I shall rest in the house of Lord to length days.* Mr. Zeldner points out that a great Hebrew poet once said that translation "is like kissing your sweetheart through a veil."

It is, in fact, amply clear that a translation procedure that does little more than handle a one-to-one correspondence of words cannot hope to be useful for problems of *literary* translation, in which style is important, and in which the problems of idiom, multiple meanings, etc., are frequent.

Even this very restricted type of translation may, however, very well have important use. Large volumes of technical material might, for example, be usefully, even if not at all elegantly, handled this way. Technical writing is unfortunately not always straightforward and simple in style; but at least the problem of multiple meaning is enormously simpler. In mathematics, to take what is probably the easiest example, one can very nearly say that each word, within the general context of a mathematical article, has one and only one meaning.

The Future of Computer Translation

The foregoing remarks about computer translation schemes which have been reported do not, however, seem to W. W. to give an appropriately hopeful indication of what the future possibilities may be. Those possibilities should doubtless be indicated by persons who have special knowledge of languages and of their comparative anatomy. But again, at the risk of being foolishly naive, it seems interesting to indicate four types of attack, on levels of increasing sophistication.

Meaning and Context

First, let us think of a way in which the problem of multiple meaning can, in principle at least, be solved. If one examines the words in a book, one at a time as through an opaque mask with a hole in it one word wide, then it is obviously impossible to determine, one at a time, the meaning of the words. "Fast" may

mean "rapid"; or it may mean "motionless"; and there is no way of telling which.

But, if one lengthens the slit in the opaque mask, until one can see not only the central word in question but also say N words on either side, then, if N is large enough one can unambiguously decide the meaning of the central word. The formal truth of this statement becomes clear when one mentions that the middle word of a whole article or a whole book is unambiguous if one has read the whole article or book, providing of course that the article or book is sufficiently well written to communicate at all.

The practical question is: "What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word?"

This is a question concerning the statistical semantic character of language which could certainly be answered, at least in some interesting and perhaps in a useful way. Clearly N varies with the type of writing in question. It may be zero for an article known to be about a specific mathematical subject. It may be very low for chemistry, physics, engineering, etc. If N were equal to 5, and the article or book in question were on some sociological subject, would there be a probability of 0.95 that the choice of meaning would be correct 98% of the time? Doubtless not: but a statement of this sort could be made, and values of N could be determined that would meet given demands.

Ambiguity, moreover, attaches primarily to nouns, verbs, and adjectives; and actually (at least so I suppose) to relatively few nouns, verbs, and adjectives. Here again is a good subject for study concerning the statistical semantic character of languages. But one can imagine using a value of N that varies from word to word, is zero for *he*, *the*, etc., and needs to be large only rather occasionally. Or would it determine unique meaning in a satisfactory fraction of cases, to examine not the $2N$ adjacent *words*, but perhaps the $2N$ adjacent *nouns?* What choice of adjacent words maximizes the probability of correct choice of meaning, and at the same time leads to a small value of N?

Thus one is led to the concept of a translation process in which, in determining meaning for a word, account is taken of the immediate ($2N$-word) context. It would hardly be practical to do this by means of a generalized dictionary which contains all possible phases $2N + 1$ words long: for the number of such phases is horrifying, even to a modern electronic computer. But it does seem likely that some reasonable way could be found of using the microcontext to settle the difficult cases of ambiguity.

Language and Logic

A more general basis for hoping that a computer could be designed which would cope with a useful part of the problem of translation is to be found in a theorem which was proved in 1943 by McCulloch and Pitts.[1] This theorem states that a robot (or a computer) constructed with regenerative loops of a certain formal character is capable of deducing any legitimate conclusion from a finite set of premises.

Now there are surely alogical elements in language (intuitive sense of style, emotional content, etc.) so that again one must be pessimistic about the problem of *literary* translation. But, insofar as written language is an expression of logical character, this theorem assures one that the problem is at least formally solvable.

Translation and Cryptography

Claude Shannon, of the Bell Telephone Laboratories, has recently published some remarkable work in the mathematical theory of communication.[2] This work all roots back to the statistical characteristics of the communication process. And it is at so basic a level of generality that it is not surprising that his theory includes the whole field of cryptography. During the war Shannon wrote a most important analysis of the whole cryptographic problem, and this work is, W. W. believes, also to appear soon, it having been declassified.

Probably only Shannon himself, at this stage, can be a good judge of the possibilities in this direction; but, as was expressed in W. W.'s original letter to Wiener, it is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?

This approach brings into the foreground an aspect of the matter that probably is absolutely basic —namely, the statistical character of the problem. "Perfect" translation is almost surely unattainable. Processes, which at stated confidence levels will produce a translation which contains only x percent "error," are almost surely attainable.

And it is one of the chief purposes of this memorandum to emphasize that *statistical semantic* studies should be undertaken, as a necessary preliminary step.

The cryptographic translation idea leads very naturally to, and is in fact a special case of, the fourth and most general suggestion: namely, that translation make deep use of language invariants.

Language and Invariants

Indeed, what seems to W. W. to be the most promising approach of all is one based on [. . .] an approach that goes so deeply into the structure of languages as to come down to the level where they exhibit common traits.

Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another, they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But, when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers.

Thus may it be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common base of human communication—the real but as yet undiscovered universal language—and then re-emerge by whatever particular route is convenient.

Such a program involves a presumably tremendous amount of work in the logical structure of languages before one would be ready for any mechanization. This must be very closely related to what Ogden and Richards have already done for English—and perhaps for French and Chinese. But it is along such general lines that it seems likely that the problem of translation can be attacked successfully. Such a program has the advantage that, whether or not it lead to a useful mechanization of the translation problem, it could not fail to shed much useful light on the general problem of communication.

Note

Editors' Note: This is the memorandum written by Warren Weaver on July 16, 1949. It is reprinted with his permission because it is a historical document for machine translation. When he sent it to some 200 of his acquaintances in various fields, it was literally the first suggestion that most had ever seen that language translation by computer techniques might be possible.

References

1. Warren B. McCulloch and Walter Pitts, *Bull. Math. Biophys.*, no. 5, pp. 115–133, 1943.

2. For a very simplified version, see "The Mathematics of Communication," by Warren Weaver, *Sci. Amer.*, vol. 181, no. 1, pp. 11–15, July 1949.

This page intentionally left blank

2
Mechanical Translation

A. D. Booth

During the summer of 1947 I first suggested that a digital computer having adequate memory facilities could perform the operations necessary to translate a text written in a foreign language (FL) into the desired language or target language (TL). There was, and is, no particular difficulty in doing this, as I hope to show in the present article; but I make no claim that a literary quality in the result of the translation is to be hoped for.

The original proposals covered only the making of a straightforward dictionary translation from the foreign language to the target language. It is convenient to start by seeing how this simple objective may be achieved on a machine whose primary purpose is the manipulation of numbers. It is necessary to assume only the most rudimentary machine functions in order to perform mechanical translation (MT):

a. The machine has a large memory.

b. The input typewriter sends data, either direct to the memory, or to a register provided with subtraction facilities, the accumulator register.

c. The machine contains a conditional transfer order which enables the machine to select between alternative courses of action according to the sign of the number held in the accumulator register.

d. The contents of the accumulator can be typed at the output.

The reader familiar with modern automatic digital computers will see that all of the above functions are present in all such computers existing, with the exception in many cases of the large memory.

How shall we represent the foreign language text in digital form? A normal teletype machine is so constructed that the depression of any key, for example that corresponding to letter A, causes the emission of a binary coded digit pattern which has a one–one correspondence to the desired character. Thus: A becomes 00011; B becomes 11001; C becomes 01110; . . . ; and Z becomes 10001. It follows that, if the keys corresponding to the letters of the foreign word are depressed, in sequence, a digital pattern will be generated which uniquely represents that word. If this pattern is regarded as a number, a dictionary translation of the foreign word can be obtained by storing the translation in that memory location which has the same number as the code of the foreign word. As an example, the Latin word *et* is coded 10000, 00001, which as a binary number is equal to 512 plus 1 or 513 and would identify memory location 513. Then in that memory location 513 we would store the translation: 00100 (d), 01100 (n), 00011 (a) corresponding to "and." The reader interested in details will notice that it is assumed that digits are shifted into the machine register, starting from the least significant (right shift), and that the inversion of order (d, n, a) is necessary for the output type to appear in the normal sequence.

It is at once obvious that this simple scheme is quite impracticable, since even in the example given, it will be seen that 1024 locations are required to deal even with the two letter words of the foreign language. For words of maximum length say 10 letters, 2^{50} locations would be needed. This would exceed even the most sanguine hopes of modern machine designers. In any case no known foreign language has anything approaching 10^{15} (2^{50}) words, so that almost all of the memory would be empty.

The difficulty is easily overcome, however. Suppose that each location (in sequence) in the memory contains a "dictionary" word (DW) having the following composition: the FL word (10 letters say) and the TL translation (40 letters say). Assume that the DWs are stored in ascending order of magnitude. Then if the FL word is subtracted from each of the DWs in turn, the result will be negative until the required entry is reached and positive thereafter. It follows that, if the conditional transfer is used to break off the sequence of subtractions at the first positive result, the remainder in the accumulator at this point will represent the target language translation. The latter may now be printed as the output.

A second obvious point is that the length of the required words (250 binary digits or *bits* in the above example) is considerable. Existing computing

machines fall short of dealing with this by a factor of five or greater. They may, however, easily be programmed to use multiple length words so that this is not an essential difficulty.

If the actual FL word is not contained in the memory, the nearest equivalent will be generated by the above process. Furthermore, since the DW, FL entry will be numerically somewhat larger than the text FL word, the output operation will generate certain nonsensical characters before the TL translation. This will indicate to the reader that an untranslatable word is present.

The preceding simple scheme is much limited by the available memory in existing (and near future) machines. But in 1948 R. H. Richens suggested to me a modification which makes mechanical translation a really practicable operation. Richens pointed out that, with certain limitations, an adequate or passable translation of a foreign language text would result from the following operation:

a. The memory contains a *stem* (or *root*) dictionary and an ending dictionary.

b. The stem dictionary consists of a relatively few entries of general semantic utility plus a vocabulary specific to the subject of the translation.

(The latter has since been called, by V. Oswald, Micro-semantics).

The method of operation is simple. First the FL word is subtracted in turn from the entries in the stem dictionary. In this way, the longest possible stem entry is found. At this point the stem translation and suitable grammatical notes are typed out. The stem is now removed from the FL word, and the remainder is compared with the entries of the ending dictionary. When coincidence is attained again, the relevant syntactic information, contained with the ending entry, is typed out.

Richens has shown that the same method can be applied to multiple words of the type encountered in, for instance, German.

As an example of this procedure consider the translation of the Latin word *amo*. This would proceed as follows:

Stem:	Trial 1:	*a*, alas
	Trial 2:	*am*, love (v) (v for "verb")
Ending:	Trial 1:	o, (1.s.p.) (for "1st person singular, present tense")

The total output would be: love (v) (1.s.p.)

Certain difficulties arise, as in the example *desider-emus* given by Richens. Here two possible translations exist: (1) *desider*, desire; *emus* (1.p.s.a.); or (2) *desid*, be idle; *eremus* (1.p.i.s.a.). Resolution could be attained by storing the word itself, together with both translations.

Again, certain words, or parts of words, are sometimes without significance, for example the *t* in the French *a-t-il*. In this case, to avoid confusing the operator, the machine probably would have to put out some encouraging symbol, such as "N" for no significance.

It has been suggested, by Prof. Erwin Reifler of the University of Washington, that semantic ambiguities could be considerably eliminated by the use of a person called a "pre-editor" who could be a native in the FL but would not necessarily know the TL at all. The duty of the pre-editor would be to replace all ambiguous words by non-ambiguous equivalents.

The foregoing brief account of mechanical translation is naturally incomplete in many respects. The act of coding a given example for a particular computer involves many points which it has been impossible to cover in a short article. This is particularly true of the *stem-ending* dictionaries, whose use requires a high degree of sophistication in the program if a good working speed is to be attained.

Some of these problems however have been actually examined on our computer APEXC at Birkbeck College, London, and the reader may be interested in the following statistics:

Time taken to translate a 1000-word message by a skilled bilingual human being:	1 hr.
Time of mechanical translation using the above technique on standard punched-card equipment:	1 hr. 54 mins.
Time of mechanical translation on APEXC using teletype output:	2 hrs. 15 mins.
Time of mechanical translation on APEXC with tabulator output:	30 mins.

It does not appear likely that with existing input-output equipment any much greater speed is possible. The translations, produced by the above methods are of course inelegant, but are easily understood by a person expert in the subject of the paper. Neither the present author nor Richens envisages the literary use of mechanical translation in the near future or even foreseeable future; but within its limitations, the method should be of great use to students and institutions confronted with the mass of published material in foreign languages which is currently appearing.

3
The Mechanical Determination of Meaning

Erwin Reifler

However one may feel about the progressive substitution of mechanical operations for the work of human beings, the fact is that the ever-increasing volume of important publications in many languages, the insufficient number of competent translators, and the time consumed in translation all justify a search for a mechanical solution to the problem of high-speed mass translation.

Whoever has had any experience with translation from one language into another will be acquainted with the polymorphic disparities characterizing every set of two languages. It is, therefore, of the greatest importance to emphasize that those who first envisaged and subsequently followed up the possibility of a mechanization of the translation process did not do so in blissful ignorance of the manifold difficulties of the task, but were well aware of the linguistic and engineering problems involved. Warren Weaver's memorandum "Translation" exhibits this awareness of the obstacles that lie in the way of a complete mechanization of the translation process.

My own first reactions to this new expansion of the empire of the machine furnish an illustration of the scepticism and sometimes rather limited ambition which characterized most of the pioneers. My research in comparative semantics,[1] my experience in translation, and my teaching of foreign languages made me at first relegate MT to the realm of the impossible. In the course of further research, however, I began to see certain limited possibilities. In a number of papers [...] I described certain language problems with which the MT linguist may be confronted and outlined some possible solutions. But I insisted that the lack of semantic distinctiveness of the conventional graphic form of languages made it unavoidable to include in the process the cooperation of a human pre-editor. However, the results of my subsequent research, which I shall outline below, showed that I was mistaken and that all human pre-editorial work, which I had previously considered a *conditio sine qua non*, could be completely mechanized.

MT Linguistics

Both the traditional linguist and the MT linguist will endeavor to observe and describe all essential aspects of the phenomenon language, including its mechanics. There are, however, the following fundamental differences between the two men. The former will, wherever possible, tend to make spoken language, the primary symbolization of language, the object of his investigation. Only when this is not accessible, or for additional information, will he turn to secondary symbolizations. He will in his analysis of a particular language ignore everything—even linguistic information—that is extraneous to the language under investigation. On the whole, he will not let himself be influenced by considerations of practicality, at least not in the first instance.

The situation is quite different for the MT linguist. It is true that both the phonic and the graphic forms of language fall within the sphere of interest of MT, but the conventional phonic symbolization of free forms is often less distinctive than their corresponding graphic representation. Striking examples are English **to, too, two,** German *her*, *hehr*, *Heer*, etc. Homophony plays an even greater role in languages like Chinese and Japanese. In such languages the historical form of writing is "symbolico-semantically" much more distinctive.[2] Since the graphic form of a language generally leaves less to be inferred from situational criteria than its spoken form, it will mostly present MT with a less formidable problem. Consequently the MT linguist will—at least at this early stage of development—mainly study language in its conventional graphic form, in which it is not homophones but only homographs that will concern him.

The MT linguist, moreover, will be mostly concerned with differences in behavior between a given pair of languages. He need not adhere strictly to the results of scientific language research. When they serve his purpose, he will consider them. But he will ignore them when an arbitrary treatment of the language material better serves his purpose.

Important examples are the *attributable universals*. These will be discussed later in connection with the mechanical correlation of certain input and output forms which, although, strictly speaking, of different meaning, can, for all practical purposes, be considered equivalents. Practicality, for the MT linguist, is a consideration of the highest order.

Furthermore, MT is concerned primarily with *meaning*, an aspect of language that has often been treated as a poor relation by linguists and referred to psychologists and philosophers. *The first concern of MT must always be the highest possible degree of source-target*[3] *semantic agreement and intelligibility*. The MT linguist, therefore, must study the languages that are to be mechanically correlated in the light of source-target semantics.

The understanding of the import of a message depends on both unsymbolized situational criteria and information supplied by the message itself. The latter is either of a nongrammatical or of a grammatical nature. It is true that one need not have studied the grammatical structure of a language or know how to label its grammatical categories in order to understand it. This does not mean, however, that a native speaker of a language, although he has never studied its grammatical structure, does not actually use grammatical criteria in the process of "comprehending" another speaker or writer in his own language. When, for example, a Chinese has "comprehended" that the utterance of another speaker in his language means something like "the dog bites the man," he has noted (1) that the word for dog precedes the word for bites, (2) that no other word intervenes between "dog" and "bites" to indicate that the dog does not carry out but undergoes the action, and (3) that the word for man follows the word for bites and hence the man undergoes the action. He will be quick to point out these (grammatical!) facts if his interlocutor, in his subsequent speech or gestures, implies that what he meant was "the man bites the dog."

Thus, apart from unsymbolized situational criteria, we depend for our "comprehension" on both nongrammatical and grammatical criteria. A complete message contains information which, together with a certain number of unsymbolized situational criteria, enables the human hearer, reader, or translator to select the intended meanings from the multiple potential meanings characterizing its constituents. But it is frequently not necessary to listen or read to the end before one is able to make the right choice. We may have to consider nothing but the nongrammatical meaning of only one individual meaningful constituent of the message. This will be so when, in all possible environments, the constituent concerned has only one nongrammatical and grammatical meaning. Examples are certain interjections like German *he* (I say) or any free symbol sequence whose boundaries coincide with those of the total message such as German *Jawohl* (yes). Or we may have to consider the nongrammatical and/or grammatical meaning of more than one meaningful constituent co-occurring in the context before we can make up our mind concerning the intended meaning of a particular constituent. Moreover, our decision about the nongrammatical meaning of such a constituent will often depend on its grammatical relation to other constituents.

Such meaningful constituents may be either free or bound minimal symbols like German *A* in *Wer A sagt, muss auch B sagen* (literally: **He who says *A* must also say *B***) or *U* in *U-Boot* (**submarine**). Or they may be either free symbol sequences like German *Luft* (**air**), or *Luftschiff* (**airship**), or bound symbol sequences like German *Luft* or *schiff* in *Luftschiff*. Or, finally, such an individual meaningful constituent may be a "clue set": that is, a group of individual free symbol sequences of the same context, one or more of which pinpoint the intended meaning of the remainder of the set.[4]

The *graphic* form of languages, as we have emphasized above, leaves less to be inferred from situation; it is more explicit than the spoken form. It is generally sufficiently explicit for intelligent, educated, or specialized readers. Nevertheless, much may have to be extracted from the context before we can arrive at a decision concerning the intended meaning, at the lowest level of free or bound minimal symbols, on the level of free or bound symbol sequences, on the level of clue sets, or on the level of higher contextual units. Both nongrammatical and grammatical meaning may be involved. Moreover the decision concerning the first may depend on a previous decision concerning the second. We are usually not aware of the complexity of the thinking process that ultimately leads to "comprehension." Like the Chinese mentioned above, we do not even realize that we have considered not only nongrammatical meanings but grammatical criteria as well. We have the impression that our determination of intended meaning is instantaneous. This impression is due to the extraordinary speed with which we comprehend. But it is, nevertheless, an illusion. Our comprehension which culminates in understanding is *progressive*.

The question that confronts MT here is whether the manifold progressive operations of the human brain can be imitated by the operation of a mechanical brain. The number and complexity of operations necessary give the impression that we are faced with a total too astronomical to be taken seriously. That is the reason why I originally took the position that a complete mechanization of the translation process will not be possible—at least for considerable time to come—and that no more than a limited MT, which speeds up certain operations of the translation process, may be attained in the near future. Mechanical processes would be inserted somewhere between the original language text and the translation product, but the cooperation of a human agent who intervenes, between the original text and the mechanical system, or between the translation product and the ultimate reader, or both, would be unavoidable.

The Problem of Editing

A correlation of different languages is only possible if they share certain aspects. All languages actually do have a number of features in common. They all depend in their spoken form on the same speech apparatus used in a more or less similar way. Agreements in features which concern the logical aspect of language are especially numerous. They are found in the grammar[5] as well as in the lexicon.[6] Of greatest importance for MT—and, of course, for translation in general, is the fact that "... as to denotation, whatever can be said in one language can doubtless be said in any other ... the difference will concern only the structure of the forms, and their connotation."[7]

The different means at the disposal of every language for symbolizing whatever can be symbolized in any other language will be outlined below. A crucial problem, however, for the success of MT in particular is the formal distinctiveness of the conventional symbols of the languages considered for MT.

Post- and Pre-editing

Let us first consider the *post-editor* and the work he would have to do. At the MT conference at the Massachusetts Institute of Technology, most participants believed that it would be possible to build a mechanical system capable of extracting a certain amount of essential grammatical information from the conventional graphic form of source language texts without the intervention of a human agent. The determination of intended nongrammatical meaning, however, most MT linguists considered to be of an order that would make mechanization impractical. The solution favored by most was to leave to a post-editor both the determination of the remaining essential grammatical information and the determination of the intended meaning appropriate to the context in all cases of multiple nongrammatical meaning. In a sense, such a system would do even more than a human translator, for it would supply the output equivalents for more than one nongrammatical meaning of ambiguous constituents of the input text. But here one cannot speak of a complete mechanization of the translation process, because it excludes from mechanization one of the most important aspects of all translation—the choice of the particular output equivalent indicated by the context. Yet this proposal is very engaging because the post-editor is required to know only the target language; no human agent familiar with the source language need appear anywhere in the translation process.

The determination of the intended nongrammatical meaning could be assigned to a *pre-editor*. He would have to be familiar with the language of the original text and, adding symbols such as diacritic marks, would increase the semantic explicitness of the conventional graphic form of the original text sufficiently to allow a mechanical system to supply the correct output equivalent in every case. In support of the pre-editor proposal, I pointed out that the task of the pre-editor would be much easier to accomplish than that of a post-editor because the former would determine intended meanings in a context completely intelligible to him. The post-editor, on the other hand, would have to do his work in an output context, necessarily containing a large number of nondistinctive free symbol sequences (words), each of which has multiple meanings. Moreover, out of each cluster of potential output equivalents, he would have to decide which equivalent fits the context, in relation to other clusters of potential output equivalents among which he often would not yet have made his choice. That is, he would often be faced with an output context that would be far from clear.

Thus I concluded at that time that the simplest form of MT would be one with pre-editing, in which the pre-editor determines the meaning indicated by each context and denotes it by special graphic symbols which he adds to the conventional written form in all instances involving multiple meaning. There is, however, another important aspect of translation that has to be taken into consideration: The determination of intended meaning depends not only on the semantic peculiarities of the source language, but on the

semantic peculiarities of the target language as well! As already mentioned, our problem is *multiple meaning in the light of source-target semantics*. If, for instance, we want to translate the English sentence, "He is an ass," into Chinese, we must discover whether the Chinese word for "ass" can be used as a contemptuous expression denoting a stupid human being. As a matter of fact, it cannot be so used, and therefore a literal translation would be completely unintelligible. Another Chinese word meaning something like "stupid" or "foolish" has to be substituted or else the English sentence has to be expressed in a completely different way according to the idiomatics of the Chinese language.

We could put the burden of the semantic interpretation on the customers of MT: that is, those who want foreign works mechanically translated. This approach is closely bound up with the problem of writing for mechanical translation. Expressed in the most general terms, writing for MT means that people desirous of machine translation of foreign language material submit the material to the MT center in a specified form, a form whose language and/or script is better suited to MT than the original form. The form specified may be either entirely different from the conventional form or merely a modification of it. Such a procedure could appreciably simplify the engineering problem and even result in a complete mechanization of the translation process proper: i.e., translation short of semantic interpretation of the original text. The mechanical correlation of the grammatical forms of the source and target languages would also be greatly simplified by regularization of the source language.

Or we could request the customers to submit texts made graphio-semantically completely explicit by the insertion of distinctive supplementary symbols. We could develop monoglot dictionaries which would explain every one of the multiple meanings of its head entries entirely in the language to which these head entries belong and in the light of the semantic peculiarities of each of the target languages involved, and which would indicate each possible meaning by a distinctive symbol. Such dictionaries could, of course, also be mechanized. The customers would then use one of these dictionaries and select from it the proper supplementary symbol to add to the text for machine processing. These procedures would give us a graphio-semantically *completely distinctive source text* which would allow a complete mechanization of the translation process in the narrower sense. But it is well to stress that we can here speak of a complete mechanization only because we have *excluded from the scope of MT the mechanization of the determination of intended meaning* in cases of multiple meaning.

If we think that the determination of intended meaning should be included in the scope of MT, we can pursue another approach, which includes a pre-editor knowing the source language concerned, and a mechanized monoglot dictionary of the type outlined above. When the pre-editor inputs the conventional graphic form of the text into the translation mechanism, the text passes first through the mechanical dictionary. Whenever no multiple meanings in terms of source-target semantics are involved, the input material is immediately translated. Otherwise a signal calls the attention of the pre-editor to the fact that multiple meanings are involved and the dictionary entry concerned appears on a screen. The pre-editor then selects the meaning required by the context and feeds it into the machine by means of the distinctive symbol that represents this meaning and is supplied by the dictionary entry. The machine then releases the section concerned for the next stage in the translation process.

All the solutions I have mentioned are feasible, but I believe they will remain academic. They are not practical. The burden on the supply side is too great; the extent of human intervention is too large; and the essential and most complicated aspect of MT, that of multiple grammatical and nongrammatical meanings, remains unmechanized. These solutions, furthermore, do not fulfill the ideal of an MT based *entirely* on the conventional form of the original text. On the contrary, this conventional form has to be replaced, modified, or supplemented. These solutions cannot help being too slow. They are, in short, still very far from the ideal of complete mechanization.

Moreover, my subsequent research results have deprived the pre-editor of all *raison d'être*. This leaves only the post-editor for the clarification of any ambiguities in the output text that are not cleared up by the translation mechanism. We have already said that a mechanization of the translation process which excludes the determination of the most suitable output equivalent among multiple meanings does not really deserve the name of MT; it can lay claim only to the mechanization of some of the steps in the translation process. But must we limit our ambition to such an incomplete solution?

No Editor

Dr. Weaver suggests in his memorandum: "It does seem likely that some reasonable way could be found

of using the micro context to settle the difficult cases of ambiguity." In one of my MIT conference papers I outlined, and in subsequent research I further developed, ideas aiming at such an ultimate elimination of the human post-editor. It is clear that a mechanical-translation system whose design provides for the determination of intended nongrammatical meaning must be very complex. But, so long as the technical requirements do not exceed the boundaries of practicality, I see no reason why such a solution should not be sought. It could extend the scope of MT beyond its present limitation to scientific publications, with the ultimate goal of creation of general-purpose translation machines, capable of translating even poems, as long as unconventional or even "bad" prose is satisfactory.

The greater mechanical complexity necessitated by the elimination of the post-editor makes economies on other levels of MT particularly welcome. The results of my research leave no doubt that an MT system can be built which abstracts all essential grammatical information from a conventional text without the necessity of human intervention, and that it is possible to *reduce substantially the number of lexical items to be coded into the mechanical memory*.

Let us summarize here what the ideal of MT would be. It would be an automatic system which, on the input side, swallows messages in their conventional graphic form and, on the delivery side, spews out these messages in one of the possible conventional forms allowed by the target language for which the system is built. Since such a system would not require any change in the conventional graphic form of the original message, of course it does not need either a pre-editor or a post-editor. The operator on the input side, consequently, does not need to know the language of the original text; he would be concerned only with feeding the text into the translation mechanism.

It is, of course, not necessary to aim so high. For the present, our ambitions need go no further than an output which, though not conventional, is, nevertheless, both substantially equivalent semantically and intelligible. Nor need we completely exclude the intervention of the human being in the MT process. We may allow a pre-editor whose activity is limited to the determination of the branch or subbranch of knowledge to which the text for translation belongs and who instructs the operator of the machine to press a special key, with the result that a mechanical memory selects only output equivalents characteristic of that branch of knowledge.

Compound Forms

As already pointed out, most participants in the first MT conference believed—though this belief could not be substantiated at the time—that it would be possible to build machines capable of abstracting most of the essential grammatical information from the conventional form of original texts. The only exception was the determination of the inner boundaries of the constituents of extemporized compounds. Here no mechanical solution seemed possible.[8] Now, if this were really impossible, if a pre-editor had to be retained for the indication of the "seam" of these unpredictable compounds, then the question arose whether the same human agent could not at the same time signalize grammatical criteria by the use of appropriate symbols, paying dividends in the form of savings in machinery and mechanical operations, and thus compensating for the disadvantages due to the necessity of human intervention.

As it turned out, there actually is a simple mechanical solution for the problem of the identification of the constituents of *all* compounds that are not "memorized" in toto, as long as the constituents appear in the memory. This solution means the complete elimination of the pre-editor from the identification process.

The earliest MT scheme, worked out in England in 1948 by Drs. R. H. Richens and A. D. Booth, already includes the mechanical dissection of complexes. [...] Their MT system was capable of remarkable feats in identification, analysis, and translation of such forms.

The mechanical dissection of complexes and their identification *via* the identification of their constituents means that practically no complex form, all of whose constituents are prolific and/or productive, needs to be coded into the mechanical memory. English examples are **sea-** and **-s** in **seaboard, seaside, seaway,** etc., and in **seas, boards, ways,** etc. Only the prolific and productive constituents need be coded. The increase in the number of mechanical operations which such an arrangement implies will be amply compensated for by a reduction in the size of the memory.

The obvious advantages of such a procedure could also be made available for compound forms. However, three difficulties have to be faced from the outset. One is that the meaning of a compound often cannot be inferred from the meanings of its components. This difficulty can be overcome by entering all such compounds in the memory. The other two difficulties seem at first to constitute insoluble problems.

The first is the so-called "*X*-factor" problem, i.e., a letter or letter sequence which could be part of the preceding as well as of the following constituent of a compound; the second, the problem of extemporized, i.e., unpredictable, compounds.[9]

Since I am very familiar with German, and since this language is not only comparatively rich in the formal indication of grammatical meaning, and thus graphio-grammatically highly distinctive, but also notorious for its abundance in extemporized compounds and, moreover, very important for MT, I first developed this solution for the substantive compounds of the German language and then tested it on other languages.

X Factor

The recognition that there are only four possible matching situations for *all* types of substantive compounds with an "*X* factor," together with the utilization of the distinction given to forms by the occurrence or absence of space (the space interval which separates free forms), plus separate memories for all essential types of graphio-grammatically distinctive forms, makes the mechanization of this aspect of MT possible. Furthermore, this solution is applicable to all languages that have the same problem.

A Russian example demonstrating the most complicated situation of a compound with an *X* factor is *rybolovu*. Not only the free form *ryb*, meaning **of fishes,** and *olovu*, meaning **to the tin,** will occur in the mechanical memory, but also the right-bound form *rybo-* and the left-bound form *-olovu*, the latter meaning **to the catcher.** This permits two dissections, namely *rybo|lovu*, meaning **to a fisherman,** and *rybolovu*, meaning **to the tin of fishes.** Only the first is the correct dissection and translation. The connective vowel *-o-* is here for MT an *X* factor. My solution makes sure that the mechanism will supply a unique and correct answer for all compounds of this nature after the third matching step.

No doubt this mechanical determination of the constituents of compounds is also applicable to French. But the compounds of this language happen to be mostly of a kind not requiring any translation. This will be explained and exemplified below.

Extemporized Compounds

Compounds like English **seashore** (substantive plus substantive), **highland** (adjective plus substantive), **afterthought** (adverb plus substantive), and **cutthroat** (verb plus substantive) need not be "memorized" because not only will their constituents all be recorded in the mechanical memory, but also their meanings can—at least in a large number of output languages—be inferred from the meaning of the equivalents of their constituents. These are all known compounds. Extemporized compounds also, although not so common as in German, turn up daily in English. Take a word like **holdability** ("Nails with more holdability"). Both **hold** and **ability** will occur as free forms in the memory. This memory can be so planned that the fact that the first is a *right-bound* and the second a *left-bound* form in **holdability** is taken into account, so that, in instances like these, only one of their possible output equivalents is selected by the mechanism. In a German output, they could be made to appear as *halt-* and *-barkeit*, respectively.

The same is true of Russian compounds like *nebosvod* (substantive plus substantive), literally **sky vault**, meaning **firmament;** *novosel* (adjective plus substantive), literally and freely **new settler;** *posleslovie* (adverb plus substantive), literally **afterword** (cf. German *Nachwort*), meaning **epilogue,** and *bezoblachnost* (preposition plus substantive), literally **without cloudiness,** meaning **cloudlessness.**

The application of this solution of the problem of the mechanical dissection of compounds must wait upon lexical research, which alone can determine the contents of the machine memory. It is necessary to determine each type of graphio-grammatically distinctive form and what constituents cannot occur as *last* constituents of compound proper names, to abstract all *X* factors and all *X*-factor forms, to collect all *known* compounds whose meaning cannot be inferred from the meanings of the output equivalents of their constituents, etc. An example for the last-mentioned problem is German *Mit|gift*, literally **with/poison,** but actually meaning **dowry.** [. . .]

The Mechanical Determination of Grammatical Meaning[10]

Experience and situational criteria often enable a human translator to grasp intuitively the semantic content of a foreign text whose grammatical problems he does not fully understand. My classroom experience with students of classical Chinese supplies ample evidence for this gift of the human brain. The human translator need therefore not adhere to any fixed se-

quence of determinative evaluation procedures. This is, however, not yet feasible in MT. Here a *hierarchy* of determinative operations is necessary, at least for the present.

The problem that faces us here is the minimum human intervention necessary or the maximum mechanization possible to make MT both feasible and practical. In order to achieve the maximum possible mechanization, we first have to isolate the individual determinative steps that the translation process requires, to determine the sequence in which the human translator takes these steps for every type of individual translation problem, and then to study these in the light of the requirements of mechanization.

A translation mechanism that embodies in its design the wherewithal for the mechanization of a human pre-editor's work has to deal with the source-target semantic problems in stages according to a hierarchical sequence. It will, within the limitations of its design, first determine the grammatical situation of each source form and then, on the basis of the grammatical situation, proceed to the determination of its intended nongrammatical meaning and the supply of the appropriate output equivalent.

At both stages it will first locate in its memory the meanings of each *source form in isolation* and then proceed to the "pinpointing" of the *intended* meaning of those source forms that in isolation have multiple meanings. This the mechanism will do through the determination of *semantic coincidences* exhibited by syntactically correlated co-occurrences in the input text. This determination of semantic coincidences depends on the information a specially planned mechanical memory supplies after a source form has been matched with its memory equivalent. In the memory, the equivalent of each source form will be accompanied by code signals, each signal indicative of one of the possible grammatical or nongrammatical meanings of the source form concerned. A meaning which a source form has only when co-occurring with one or more other syntactically correlated source forms will be indicated *by the same code signal* accompanying the memory equivalents of each of these source forms.

If there are multiple grammatical meanings, the translation mechanism extracts the intended meaning by determining the grammatical code signals in which two or more syntactically correlated source forms coincide. For example, *den* in German can be either accusative masculine singular or dative plural in isolation. *Männern* can only be dative plural. The grammatical meaning in which both *den* and *Männern* in *den Männern* (to the men) coincide is that of dative plural. The co-occurrence of these two forms shows that here *den* can only be dative plural.

From among multiple nongrammatical meanings the translation mechanism will extract the intended meaning by determining the nongrammatical meaning in which two or more syntactically correlated source forms coincide. For example, in *er bestand die Prüfung* (he passed the examination), the memory equivalent of *bestand* will be accompanied by a number of distinctive code signals, each indicative of one of its multiple nongrammatical meanings. One of these code signals will be identical with a code signal accompanying the memory equivalents of all substantives which, as objects of *bestand*, "pinpoint" the intended meaning of the latter as one best translated by English "passed." In every instance, the determination of the intended nongrammatical meaning will be made in a section of the input text which the translation mechanism has *previously* clarified grammatically within the limitations of its design. The solution of residual semantic ambiguities will be left to the intuition and the specialized knowledge of the ultimate reader of the output text.

In an earlier report,[11] I indicated that, since substantives also occur as first words after a final punctuation mark, and thus lose the graphic distinctiveness of the initial capital letter, certain measures have to be taken in order to make sure that *all* substantives reach their matching center via the shortest possible route. [...]

Another problem, to which the cooperation of a human pre-editor offers the simplest and at present the only practical solution, is those substantives which not only are members of the general vocabulary but also occur as proper names. Therefore, instead of being supplied to the output in an untranslated form, they would be translated. Examples are the German family names *Bauer* and *Gerber* whose semantic equivalents in English are **farmer** and **tanner.**[12] [...]

Once an input form has reached its particular memory and been matched against its memory equivalent, the code signals accompanying the latter will supply all information necessary for the remaining steps in the MT process, so that no human intervention at all will be necessary on the input side. Further on I shall describe a sorting procedure that makes possible a mechanization of the pre-editorial

determination of all form classes. But first a few definitions are necessary.

Alphabetic and Nonalphabetic Constituents of Source Compounds

Punctuation marks, including links (i.e., the hyphen), separators (i.e., the comma), and also the space that is indicated by the action of the typewriter space bar, are considered integral parts of the meaningful letter sequences they precede or follow. *They simultaneously belong to the preceding as well as to the following letter sequence*. Space not indicated by the action of the space bar—that is, the space preceding the very first form of the text and that following the final punctuation mark of the text—is here ignored.

Punctuation marks and space-bar spaces, together with the letter sequence they precede or follow, form *source compounds*. We distinguish between alphabetic and nonalphabetic signals as constituents of source compounds.

Free and Bound Forms

Only a complete paragraph of a source text can properly be called a "free form," and it then includes *all signals* from the initial letter of the paragraph to the final punctuation mark after its last form. It is nevertheless important to distinguish between free and bound forms. We shall therefore use the terms free and bound in the *narrow* sense.

Free Forms The term "free form" will be applied to a meaningful alphabetic signal sequence separated from other such signal sequences by nonalphabetic signals or characterized by a capital initial letter and followed by a nonalphabetic signal (i.e., the case of the very first form of a source text).

Bound Forms The term "bound form" will be applied to a meaningful alphabetic signal sequence that has an initial small letter and is not separated from other alphabetic signal sequences by nonalphabetic signals on either side (i.e., *fahrts* in *Schiffahrtsgesellschaft*).

Half-Bound Forms The term "half-bound form" applies to a meaningful alphabetic signal sequence that is separated from another alphabetic signal sequence by a nonalphabetic signal on one side. (In the very first form of a paragraph, the initial capital is indicative of the freedom on the left side.) Consequently, we distinguish (a) left-bound forms (i.e., *gesellschaft* in *Schiffahrtsgesellschaft*) and (b) right-bound forms (i.e., *Schif* in *Schiffahrtsgesellschaft*).

The "Pinpointing" of Composite Intended Meanings

For the purposes of MT, it is convenient to distinguish between two types of meanings: monogenetic and polygenetic.

Monogenetic Meaning An intended grammatical or nongrammatical meaning completely inferrable from a single free form will be called a *monogenetic meaning*. For example, the words *wegen* (because of), *bedürftig* (in need of), *bedarfst* (thou art in need of), and *wir* (we), in *wegen dieser Schüler* (because of these pupils), *dieser Hilfe bedürftig* (in need of this help), *dieser Hilfe bedarfst* (... art in need of this help), and *wir lieben* (we love) are grammatically perfectly distinctive even in isolation.

Polygenetic Meaning Isolated free forms often have multiple grammatical or nongrammatical meanings. The "pinpointing" of the intended meaning then depends mostly on supplementary information supplied by certain co-occurrent free forms. The latter may themselves be grammatically or nongrammatically ambiguous in isolation. The "pinpointing" is thus frequently reciprocal. A grammatical or nongrammatical intended meaning inferable only from a consideration of the complementary information supplied by more than one free form will be called a *polygenetic meaning*: for example, the words *dieser* (this, of this, to this, of these), *Schüler* (pupil, pupils, of pupils) and *Hilfe* (help) in the examples above. In isolation they have multiple grammatical meanings. *Dieser* is either singular masculine nominative, singular feminine genitive or dative, or plural genitive; *Schüler* is either singular nominative, dative or accusative, or plural nominative, genitive, or accusative. But the co-occurrence of *dieser* and *Schüler* narrows down the four alternatives of *dieser* and the six alternatives of *Schüler* to the only *two* alternative possibilities of either the nominative singular or the genitive plural. The semantic reduction is here reciprocal. The co-occurrence of the semantically monogenetic *wegen*, a preposition governing *only* the genitive, then "pinpoints" the grammatical meaning of both *dieser* and *Schüler* to the genitive plural. *Mutatis mutandis*, the same holds true for the other free forms in the examples given above. The co-occurrence of *dieser* and *Hilfe* narrows down the four alternatives of *dieser* and the four alternatives of *Hilfe* (either

nominative, genitive, dative, or accusative singular) to the *two* alternatives of either genitive or dative singular. The co-occurrence of the semantically monogenetic *bedürftig* or *bedarfst*, both governing the genitive, then "pinpoints" both *dieser* and *Hilfe* to the genitive singular. The co-occurrence of *wir* pinpoints the incident meaning of *lieben* to the first person plural of the present tense of the verb *lieben*.

With forms with monogenetic meanings, apart from considerations of access time, there would be no need for a multiplicity of MT memories. The situation is different with polygenetic meanings. Whenever the first matching action has resulted in the supply of multiple alternative semantic information, it is necessary to delay the final matching and translation of an input form until the first matching of another co-occurrent input form has yielded supplementary information. If the memory equivalents of *all* source forms were stored in a *single* large memory without subdivisions, we would in such instances have to release the memory for the first matching of the "pinpointer" before the "pinpointee" had been completely clarified, then make it available to the "pinpointee" again for the determination of the latter's intended meaning, and then, frequently, turn it over for the second time to the "pinpointer"; that is, a single memory would require complicated routines and a relatively long time whenever the pinpointing is reciprocal. It is clear that it would be more convenient if in such instances the memory equivalents of "pinpointee" and "pinpointer" were stored in different submemories.

"Pinpointees" and "pinpointers" mostly belong to different form classes. Sometimes, however, they belong to the same form class as, for example, *Nachricht* and *Inhalt in dieser Nachricht Inhalt* (the contents of this message), where both are substantives, and *musst* and *essen* in *du musst essen* (you must eat), where both *musst* and *essen* are verbs. As a matter of fact, this phenomenon occurs in German only in the form classes of substantives and verbs. But, as we shall see below, for MT purposes it is preferable to ignore the superclass of *verbs* and to consider only the subclasses of *principal verbs* and *auxiliary verbs*. Thus this phenomenon affects only the *capital* memory in German.

If the "pinpointer" precedes the "pinpointee," there is no problem because then nonsubstantive forms are bound to intervene (i.e., *Wein der ersten Ernte*, wine of the first harvest); the "pinpointee" is then in fact a nonsubstantive form (*der* in our example). However, we are faced with a problem if the substantive "pinpointer" immediately follows the substantive "pinpointee" (i.e., *dieser Nachricht Inhalt*, the contents of this message), for then the capital memory has to be released for the matching of the "pinpointer" (*Inhalt*) before the "pinpointee" (*Nachricht*) has been "pinpointed." Here the fact that German substantives *acting as "pinpointers" of other preceding subtantives are always written with an initial capital* (not all German substantives have a capital initial!) comes to our assistance: The substantive "pinpointee" is "pinpointed" the moment the capital initial of the following substantive "pinpointer" has been fed in, and the capital memory can be released for the matching of the following "pinpointing" substantive the moment its capital initial has been fed in. The matching mechanism can be designed to make such a release possible after the feeding of a capital initial.

If a substantive "pinpointer" that follows its substantive "pinpointee" is separated from the latter by one or more nonsubstantive forms (i.e., *dieser Nachricht tiefer und schwerwiegender Sinn*, the deep and grave meaning of this message), then, properly speaking, the nonsubstantive form immediately following the "pinpointee" is the "pinpointer," and no problem exists.

Two Groups of Form Classes

[...] Now the form classes of all languages may be divided, for MT purposes, into (a) those with a very large membership and (b) those with a comparatively very small membership. In German we distinguish:

A. Form Classes with a Very Large Membership
1. *Paradigmatic form classes:*
   (a) Substantives.
   (b) Attributive adjectives, except the invariable forms derived from certain types of substantives by means of the suffix *-er*.
   (c) Principal verbs, except those mentioned under *B1c* below.
2. *Nonparadigmatic form classes:*
   (a) Invariable attributive adjectives derived from certain types of substantives by means of the suffix *-er* (cf. *1b*).
   (b) Predicative adjectives.
   (c) Adverbs of adjectival origin.
   (d) Cardinal numbers, except the numbers *zwei* and *drei* (cf. *B1e* below).

B. Form Classes with a Comparatively Very Small Membership
1. *Paradigmatic form classes:*
   (a) *der* words.

(b) *ein* words.
(c) The auxiliary verbs of tense *haben*, *sein*, *werden*; the auxiliary verbs of mood; the principal verbs *lehren*, *heissen* (in the sense **to command**), and, in certain instances *lernen*, *finden*, *machen*, and *lassen*.
(d) Personal pronouns.
(e) The cardinal numbers *zwei* and *drei*.

2. *Nonparadigmatic form classes:*
(a) Adverbs of nonadjectival origin.
(b) Prepositions.
(c) Conjunctions.
(d) Interjections.

With the exception of the form class of interjections, all these form classes are of consequence for the "pinpointing" of intended meaning in the case of polygenetic grammatical meanings. Therefore, wherever the distinction between "pinpointee" and "pinpointer" coincides with a distinction of form class, it is preferable to assign them to different memory compartments.

Operational Form Classes

As for the operations involved in the "pinpointing" of the intended meaning in the case of polygenetic meanings, it is convenient to set up *operational form classes* as distinguished from form classes in the traditional sense. With the sole exception of the operational form class of substantives, an operational form class includes all forms of a language which with respect to the "pinpointing" of intended meaning are mutually neutral. In other words, the members of one operational form class can serve as "pinpointee" or "pinpointer" for members of another operational form class, but never for members of their own class.

The membership of an operational form class may sometimes be identical with that of a form class in the traditional sense, as, for example, the prepositions. An operational form class may, on the other hand, include several traditional form classes, as, for instance, determinative noun qualifiers (see below); or it may include only part of a traditional form class, as, for example, the operational form class that embraces all cardinal numbers except *zwei* and *drei*. Finally, several operational form classes may constitute an operational "super form class" as, for example, the three operational form classes: auxiliary verbs of tense, mood, and the principal verbs *lehren*, *heissen*, *hören*, *sehen*, *fühlen*, *lernen*, *finden*, *machen*, and *lassen* (see below).

The Small Operational Form Classes

A survey of the membership of the comparatively small operational form classes reveals the following:

The Operational Form Class of Determinative "Pro-Adjectives" (184 forms) German adjectives after *der* and *ein* words mostly exhibit the characteristics of the so-called weak declension. In this declension they have only two distinctive paradigmatic forms: one for the nominative singular of all genders and for the accusative singular of the feminine and neuter, the other for all other cases. This makes them rather undistinctive from the grammato-semantical point of view. On the other hand, the number of different paradigmatic forms of *der* and *ein* words preceding such adjective forms is much larger, and thus these words are grammato-semantically much more distinctive. They are, as a matter of fact, very important "pinpointing" factors in the determination of intended grammatical meaning of substantives and of adjectival qualifiers other than *der* or *ein* words. On the other hand, no *der* or *ein* word ever functions as a "pinpointing" factor for another *der* or *ein* word. A number of other adjectival noun qualifiers often exhibit an analogous behavior and have the same effect on adjectives following them. The same holds true for the definite article in all cases and for the indefinite article in most cases. With the exception of *der* and *ein* themselves, all such forms are called *determinative adjectives*. In the following, we shall combine the definite and indefinite articles and the determinative adjectives into a class of "pro-adjectives" since the operational behavior of all these forms justifies the creation of a single operational form class of determinative "pro-adjectives." This operational form class includes all distinctive paradigmatic forms of *ander*, *all*, *beide*, *der*, *derjenige*, *derselbe*, *dein*, *Dein*, *dieser*, *ein*, *einige*, *einzeln*, *etlich*, *etwelch*, *euer*, *Euer*, *ihr*, *Ihr*, *jeder*, *jedweder*, *jeglicher*, *jener*, *kein*, *mancher*, *mehrere*, *mein*, *sämtlich*, *sein*, *Sein*, *solche*, *unser*, *viel*, *wenig*, *welch*.

It is true that the German adjective does not take the endings of the weak declension after all forms of these determinative qualifiers. After the indefinite article and the possessive pronouns, the adjective has the strong endings in the masculine and neuter nominative and in the neuter accusative singular, and the weak endings in all other cases. After *andere*, *einige*, *einzelne*, *etliche*, *etwelche*, *manche*, *mehrere*, *sämtliche*, *solche*, *viele*, *wenige* the adjective has the strong declension in the nominative and accusative, whereas, in the genitive and dative it is sometimes strongly,

sometimes weakly, inflected. After *alle*, *beide* and *welche*, the adjective regularly has the weak forms. But, for all practical MT purposes, it is convenient to include all forms in the operational form class of determinative "pro-adjectives," not only those after which the adjective follows the weak declension, but also those after which it has strong inflections—that is, all uninflected forms such as *all*, *ein*, *kein*, *manch*, *solch*, *viel*, *wenig*. We include, furthermore, the cardinal numbers *zwei* and *drei* and all their inflected forms. These numbers remain uninflected after a *der* word and before an attributive adjective, but alone, or unqualified before a substantive, they exhibit either the uninflected forms or the inflected forms *zweier* and *dreier* in the genitive and *zweien* and *dreien* in the dative. This brings the membership of this operational form class of determinative "pro-adjectives" to 184.

The Operational Form Class of "Pro-substantives" (46 forms) The operational form class of "pro-substantives" embraces the personal pronouns, the reflexive pronoun *sich*, the reciprocal pronoun *einander*, the forms *dessen*, *deren*, *derer*, *denen* of the pronouns *der*, *die*, *das*, and all distinctive forms of *wer* and *was* and of *man*, *jedermann*, *jemand*, *niemand*, *etwas*, *nichts*. We include all personal pronoun forms which in all positions have an initial capital, but exclude all forms shared by both the personal and the possessive pronouns (i.e., *mein*, *meiner*, *dein*, *deiner*, *sein*, *seiner*, *unser*, *euer*, *ihr*, and *ihrer*, which are included in the operational form class of determinative "pro-adjectives" above). [. . .]

The Operational Form Class of Prepositions (71 forms) This form class coincides with the traditional form class of prepositions. It includes the so-called pseudo-prepositions *angesichts*, *behufs*, *betreffs*, *bezüglich*, *namens*, *seitens*, *inmitten*, *unbeschadet*, *rücksichtlich*, *hinsichtlich*. [. . .]

The Operational Form Class of Verbs Always or Sometimes Requiring a Predicate Complement (261 forms) This operational form class comprises auxiliary and principal verbs (a) that always or sometimes require a predicate complement; for example *ist* (*ist gut*, *ist gegangen*, *ist geschlagen*, *ist ein Mann*); *hat* (*hat gesehen*), *wird* (*wird gross*, *wird König*, *wird geschlagen*, *wird kommen*); *kann*, *lehrt*, *hört*, *lernt*, *lässt* (i.e. with *singen*), and (b) that are important factors in the "pinpointing" of the intended grammatical meaning of the predicate in which they occur. Because of the necessity of placing potential "pinpointees" and "pinpointers" in separate operational form classes, this form class has to be divided into the following three operational "sub-form classes":

1. The operational sub-form class of the auxiliary verbs of tense: *haben*, *sein werden*, with a membership of 56.

2. The operational sub-form class of the auxiliary verbs of mood: *können*, *dürfen*, *mögen*, *sollen*, *wollen*, *müssen*, with a membership of 88.

3. The operational sub-form class of the principal verbs: *lehren*, *heissen*, *hören*, *sehen*, *fühlen*, *lernen*, *finden*, *machen*, *lassen*, with a membership of 117.

This makes the total membership of the "super-form class" 261.

The Operational Form Class of Separated Verb Prefixes (circa 200 forms) Separated verb prefixes were originally either prepositions (i.e., *an-*, *aus-*, *bei-*, *nach-*, *vor-*, *zu-*, etc.), adverbs (i.e., *fort-*, *nieder-*, *weg-*, etc.), substantives (i.e., *heim-*, *teil-*, *statt-*, etc.), or adjectives (i.e., *frei-*, *fest-*, *still-*, etc.) and mostly still occur also as members of these form classes. However, they are graphio-semantically completely distinctive from their prototypes in these form classes and are comparatively few in number (see below under "Operational Form-Class Filter System" for a description of the distinguishing features). They are, moreover, important factors in the "pinpointing" of the composite nongrammatical meaning of the finite verbs whose separated prefixes they are. These facts justify the creation of an operational form class of separated verb prefixes. The total membership of this operational form class has to be established by a lexical count, but it will hardly amount to more than 200.

The Operational Form Class of Adverbs of Other Than Adjectival or Numeral Origin (circa 300) The total membership of this operational form class has to be determined by a lexical count. The most common members number about 300.

The Operational Form Class of Conjunctions (circa 90) Excluding forms shared with the operational form class of adverbs of nonadjectival and non-numeral origin, the operational form class of conjunctions has about 90 members.

The Operational Form Class of Interjections (circa 20) The total membership of this operational form class has to be established through a lexical count. The number of its most common members is about 20.

This survey shows that the total membership of the comparatively small operational form classes is about 1,172. Considering additional members which a lexical research may reveal, we may safely estimate the total membership to be *less than 2,000*.

Memory Systems

If magnetic drums are used for memory storage, we may distinguish between large-drum memories and small-drum memories. A number of each will be needed. Limited membership makes the above group of operational form classes eligible for entry into the small-drum memory system. Operational form classes whose membership is large will require large-drum memories.

Large-Drum System In the large-drum system we must distinguish further between drum *units* and their constituent *submemories*. The number of the latter depends on the requirements of second, third, etc., level "pinpointing" processes. Since the present chapter is limited to an outline of the requirements of the *first-level* "pinpointing" procedure, the description of the submemories, which requires a detailed discussion of many semantic problems, will be presented in a separate paper.

For the large operational form classes, a total of four large-drum units will be needed, containing the following memories:

1. *The capital memory* for substantives and substantive constituents.

2. *The attributive adjective memory*, including the cardinal numerals, except the few included in the memory of determinative "pro-adjectives."

3. *The principal verb memory*, excluding the few verbs included in the memory of verbs always or sometimes requiring a predicate complement.

4. *The predicate adjective memory*, including all adverbs of adjectival and numeral origin.

Small-Drum System In the small-drum system, an individual memory will be assigned to each of the operational form classes with a comparatively small membership. The ten small operational form classes discussed above may be entered into ten small drums, but it may be desirable with German to subdivide the class of prepositions into six sub-form classes, according to whether they require their complements in the following cases: genitive, dative, accusative, genitive and dative, genitive and accusative, or dative and accusative. This would increase the number of small drums to 15. With the smaller number of drums, a greater complexity of equipment and circuitry would be necessary to handle the prepositions. It is, thus, a matter of engineering economics whether 10 or 15 small-drum memories are used.

We therefore have a choice between 4 large drums and 10 small drums minimum, or 4 large drums and 15 small drums maximum. If, however, a reduction in equipment is more advantageous than a decrease in access time, coding and MT operations resulting from (a) the categorization of the prepositions by cases governed and (b) the assignment of an individual small-drum memory to the operational form class of interjections, then only a total of 13 drum memories would be required: namely, 4 in the large-drum system, and 9 in the small-drum system. The interjections are then best entered into the small-drum memory of prosubstantives. But the programming must include provisions for distinguishing, in a *second-level* "pinpointing" procedure, some of the interjections from their homographic doubles in the *capital* memory and in the *predicative adjective* memory: for example, *Ei!* meaning **`Ah!`** or **`Indeed!,`** and *Ei* meaning **`egg;`** *weh!* meaning woe!, and *weh* meaning **`sore.`**

Memory Sections Access time is insignificant with forms of highest frequency, that is, those belonging to the small operational form classes, but it plays an important role with the large operational form classes all of whose members are low-frequency forms. Both types of forms are easily sorted mechanically since one is identifiable in the small-drum system and the other is not. With the large operational form classes, a substantial decrease in access time may be brought about by the following arrangement and procedure. The memory equivalents of all low-frequency forms may be grouped according to the number of their component alphabetic and/or nonalphabetic minimal symbols (i.e., single letters, certain punctuation marks, bar-space stimuli) and each group assigned to a special *memory* section of each submemory of every drum unit. Thus memory equivalents with five minimal symbols would be in the five-symbol section, those with six minimal symbols in the six-symbol section, and so forth. An example is the German right-bound adjective _warm_ (as in *warmblütig*, **`warm-blooded`**) which with the left bar-space stimulus (symbolized in the example by the underlined empty space on the left), but without the right-hand one, has five minimal symbols. Another example is the

left-bound substantive nahme (as in *Teilnahme*, **participation**) which, including the right-bar space stimulus (symbolized by the underlined empty space on the right), but lacking the left one, has six minimal symbols. Each submemory of a drum unit would then consist of a number of memory sections arranged in the sequence of the minimal-symbol numbers characterizing every one of their head entries. Within each section the order would be alphabetical. The matching mechanism counts the minimal-symbol number of every form not identified in the small-drum system and thus determines the memory sections of the large-drum system which are likely to contain the memory equivalent of the form concerned. Such forms, therefore, need not be compared with the thousands of memory equivalents in the large-drum system, but only with the much smaller number of those that have the same number of minimal symbols as the input form in question. The limitations in coding space imposed by the equipment hitherto used for memory devices in any case make an extensive subdivision of the large-form class memories imperative. A subdivision based on an operational grammatical analysis for MT and resulting in a decrease in access time is obviously preferable to any other. Thus the criterion for the type and number of memory equivalents to be coded on each drum should, I believe, be essential grammatical meaning and *number of minimal symbols*.

On the other hand, recent developments in memory devices offer an extremely economical solution to the memory-storage problem posed by MT. Rotating drums cease here to be the carriers of the memory and become carriers of the reading heads. The memory equivalents are coded on a broad magnetic tape which may have any length desired or required. The most welcome consequence of this reversal of the functions of the mechanical parts involved is that only *one* rotating drum is necessary, irrespective of the size of the memory.[13] It is, of course, very simple to divide the magnetic tape into any number of memory sections.

Operational Form-Class Filter System

The conventional "bar space" (space produced by the action of the typewriter space bar) after final nonalphabetic signals (i.e., period, exclamation mark, colon, etc.) is double that after *nonfinal nonalphabetic signals* (i.e., comma, semicolon, period after an abbreviation, etc.) and that between two meaningful *alphabetic* signal sequences, whether or not they are separated by a *nonfinal nonalphabetic signal*. These two kinds of space we shall call "the double bar space" and "the single bar space." Their graphic distinctiveness makes it possible to set up a fixed mechanical procedure for the determination of the operational form class to which each source form belongs. For the German language this procedure is the following:

Step 1 All free source forms with an initial capital preceded by the single bar space (i.e., free initial capital forms in other than first positions) are immediately directed to the capital memory. In other words, the feeding in of a single bar-space signal and an initial capital letter brings the capital memory into play. For the further sorting procedure, separating "pro-substantives" and "pro-adjectives" with initial capitals from substantives, see step 7 below.

Step 2 The input of the initial letter of all other free forms, including also initial capital forms in the first position, activates the small-drum system. Such forms are first compared with the head entries in the small-drum system.

Step 3 All source forms which are members of the small operational form classes are identified and processed in the small-drum system memories. The identification and processing of such forms in the small-drum system starts the moment the final signal (i.e., the space or punctuation signal) of the signal sequence concerned has been fed in.

Step 4 The moment a signal has been fed in which occurs in a sequence position not existing in the small-drum system, the latter is disconnected and the large-drum system is connected.

Step 5 Forms thus rejected by the small-drum system are first directed to the *capital memory*.

Step 6 All forms, free and bound, identified in the capital memory are processed there. *Free* source forms rejected by the capital memory are, in a fixed sequence, redirected to the other memories.

Step 7 They are first directed to the attributive adjective memory. Here all the remaining attributive adjectives, including those that have capital initials in all positions, are identified and processed.

Step 8 Of the forms not identified in the memory of attributive adjectives, the pronominal forms, which have an initial capital and are preceded by the single bar space, are redirected to the *small-drum system* where they are identified and processed (they belong

to the operational form classes of "pro-adjectives" and "pro-substantives").

Step 9 All other free source forms rejected by the attributive adjective memory are directed to the principal verb memory. Here all principal verb forms not previously identified in the small-drum system are identified and processed. The memory equivalents of principal verb forms which may be followed by a separated prefix are accompanied by a distinctive code signal symbolizing this potentiality. The moment the matching mechanism has "sensed" such a code signal, it detains the principal verb memory at the position of the memory equivalent concerned until the separated prefix has been identified in the *small-drum system memory*. Then the principal verb is processed in consideration of the information supplied by the memory equivalent of the separated prefix. The "pinpointing" of the composite *nongrammatical* meaning proceeds on the lines described above. I may add here that separated prefixes are made graphio-semantically completely distinctive by the fact that they are immediately followed either by a punctuation mark or by certain conjunctions (i.e., *und* or *oder*) and are preceded by a finite verb form whose memory equivalent is marked by a distinctive code signal. (Note that there is more than one verb memory. For example, *sein* in *fortsein* belongs to the operational form class of verbs always or sometimes requiring a predicate complement. Its memory equivalent is in the small-drum system, whereas the memory equivalent of *gehen* in *fortgehen* is in the *large-drum* system principal verb memory unit). If no separated prefix is found in the position assigned to it by the German language (that is, immediately preceding a punctuation signal or certain conjunctions), then the small-drum system is disconnected and the finite verb form is processed, disregarding the code signal indicative of a potential separated prefix.

Step 10 All forms rejected by the principal verb memory are redirected to the memory for predicate adjectives and adverbs of adjectival and numeral origin, where such adjectival and adverbial forms are identified and processed.

Step 11 All source forms not identified in any of the memories are forwarded to the output side in their original symbols.

Conclusion

The mechanical determination of intended grammatical meanings is, of course, not an aim in itself, but only a means to an aim. We are here not interested in the creation of machines for the sole purpose of grammatical analysis of input texts. Our aim is to provide the wherewithal for an MT product of high source-target semantic accuracy and output intelligibility without the intervention of a human agent.

In order to attain this goal, a further elaboration of details is necessary to simplify the mechanical synthesis of the intended meaning of "pinpointees" and "pinpointers" which, like those of German compound verbs with a separated prefix, are set off by a space and/or other free signal sequences.[14] But the operational form-class filtering system described here, together with the mechanical determination of the constituents of substantive compounds not separated by nonalphabetic signals which I have outlined, amply demonstrate the feasibility of a mechanization of the work of a human *pre*-editor whose intervention had previously been held to be necessary. Nor does it appear from present indications that a human *post*-editor will be necessary.

Notes

1. A branch of general linguistics concerned with the collection, comparison, and evaluation of instances of independent analogous semantic change found in words of unrelated languages. Cf. Erwin Reifler, "La 'Fission de l'atome' en sinologie à l'aide de la sémantique comparative," *Bull. Univ. l'Aurore*, Shanghai, 1949, and "Linguistic Analysis, Meaning and Comparative Semantics," *Lingua*, Haarlem, Holland, 1953.

2. In certain languages, the situation is reversed. In Hebrew, for example, the most common graphic form rarely indicates the vowels and thus is symbolically less distinctive than Hebrew speech.

3. *Source* and *target* denote the conventional form of the languages to be translated from and into, respectively; *input* and *output* refer to the text fed into or out of the machine.

4. This concept of a "clue set" was first developed in my "Studies in Mechanical Translation, no. 5," published in *MT*, vol. 1, no. 2, pp. 28–9, Aug. 1954.

5. "Pour constituer une grammaire générale dont les lois soient conformes à la réalité, la première tâche paraît donc être de dresser un répertoire de tous les faits de grammaire observés dans toutes les langues. Seule une enquête complète poursuivie méthodiquement fera connaître le caractère propre, l'étendue, la fréquence de chacune des catégories possibles de l'entendement humain. Un système à priori bâti par raisonnement abstrait sans base solide dans le réel, devient rapidement caduc. L'intuition est dangereuse quand elle se substitue aux enseignements de l'expérience" (Joseph Vendryes, "La Comparaison en linguistique," *Bull. Soc. Linguistique*, *Paris*, 1945).

6. Vendryes, *op. cit.*, pp. 16–17.

7. Leonard Bloomfield, *Language*, New York, p. 278, 1933.

8. Victor A. Oswald, *Microsemantics* (mimeographed): "We have no mechanical process by which this could be accomplished, but an intelligent … pre-editor could indicate the dissection for any sort of context"!

9. A detailed description of "The Mechanical Dissection of Known and Unpredictable Compounds" has been submitted for publication in MT.

10. First treated in my "Studies in Mechanical Translation, no. 8."

11. "Studies in Mechanical Translation, no. 7."

12. As a result of my research during the summer of 1953, made possible by a grant from the Rockefeller Foundation, I can state that there is a very simple mechanical solution for a large number of these ambiguous cases. The results will be published in a separate paper.

13. For further details, cf. "The Tapedrum, a New Brush Rapid-Access High Capacity Magnetic Memory," *Bull. 4310-1-54*, Clevite-Brush Development Co., Cleveland, Ohio.

14. The basic ideas underlying this refinement will be outlined in a forthcoming paper, "The Elimination of the Human *Post*-Editor."

This page intentionally left blank

This page intentionally left blank

(noun) "step, pass, passage, way, strait, thread pitch, precedence," and *est* (present 3rd singular verb) "is," (noun) "east." The probability of selecting the correct meaning can be increased by programming such as the following for *pas*: "If preceded by a verb or adverb, then choose 'not'; if preceded by an article or adjective, choose 'step', etc." Experiment shows this rule (and a similar one for *est*) has a confidence coefficient of .99 of giving the correct translation.

A more complicated type arises when a word has several meanings as the same part of speech. Here we can only look forward to an approach such as that suggested by Yngve, using the syntax rather than grammar. This type, of course, has by far the largest frequency of occurrence.

The formulas above use grammar (and we hope someday syntactical context) to increase the probability. The human mind uses in addition other types of clue. A fairly simple type, and hence one easily mechanized, is the association of groups or pairs of words (without regard to meaning). These are the well-known idioms and word pairs. In the system proposed the probability of correct translation of words in an idiom is increased almost to unity by actually storing the whole idiom (in all its inflected forms) in the store. The search logic of the machine is peculiar in that words, or word groups are arranged in decreasing order on each "page," so that the longest semantic units are examined first. Hence no time is lost in the search procedure. Available capacity is the only criterion for acceptance of a word group for entry in the dictionary. The probability that certain word groups are idiomatic is so high that one can afford to enter them in the dictionary.

In principle, the same solution applies to word pairs. For example, *état* has several meanings, but usually *état gazeux* means "gaseous state." Can one afford to put this word pair in the dictionary? Only experiment, with a machine, can determine the probabilities of occurrence of technical word pairs. Naturally, there will be room for some, and not for others. The exceptions lie in the same ground that we cannot approach with grammatical clues, but which may be solvable with the syntactical approach, although at the moment the amount of information which would have to be stored seems to be much too large.

The choice of multiple meaning like "dream/consider" (Fr. *songe*) is not of first importance. The ultimate reader can make his own choice easily. The multiple meaning merely clutters the output text.

The choice of multiple meaning of the so-called unspecified words like *de* (12 meanings), *que* (33 meanings) is much more important for understanding a sentence. The amount of cluttering of the output text by printing all the multiple meanings is very great, not only because of the large number of meanings for these words but also because of their frequent occurrence. Booth and Richens proposed printing only the symbol "z" to indicate an unspecified word; others have proposed leaving the word untranslated, and others have proposed always giving the most common translation. These seriously detract from the understandability. At the other extreme, one could give all the meanings. In the case of unspecified words, the reader can rarely choose the correct one, so he is given very little additional information at the expense of reducing the ease of reading.

The stochastic approach of printing only the most probable permits the best effort in making sense and prints only one word, so it is easy to read. What is the probability of successful translation?

Let us look at a few unspecified French words. Large samples of *de* have been examined. In 68% of the cases "of" would be correct; in 10% of the cases *de* would have been part of a common idiom in the store, and hence correct; in 6% of the cases it would have been associated as "de l'," "de la" which are treated as common word pairs, and hence in the store. In another 6% of the cases it would have been correctly translated by the rule sent to the data processor from the store: "If followed by an infinitive verb, translate as 'to'." Another 2% would have been obtained by a more elaborate rule: "If followed by adverbs and a verb, then 'to'." The single example of *de le* + *verb* probably would not have been programmed or stored.

There remain then 8–10% of the cases where "in, on, from" should not be translated at all. In some of the cases "of" could have been understandable, just as in the title of this paper "Stochastic Methods of Mechanical Translation" and "Stochastic Methods in Mechanical Translation" are equivalent. Further study, of course, may reveal some other rules to reduce this incorrect percentage.

Not all unspecified words can be guessed with as high a probability, but the bad cases seem more subject to programming.

In summary, we believe that this type of attack can be quite successful, but only after a large-scale study *with the aid of the mechanical translation machine itself*.

5
A Framework for Syntactic Translation

Victor H. Yngve

Introduction

The current MIT approach to mechanical translation is aimed at providing routines intrinsically capable of producing correct and accurate translation. We are attempting to go beyond simple word-for-word translation; beyond translation using empirical, *ad hoc*, or pragmatic syntactic routines. The concept of full syntactic translation has emerged: translation based on a thorough understanding of linguistic structures, their equivalences, and meanings.

The Problems

The difficulties associated with word-for-word translation were appreciated from the very beginning, at least in outline form. Warren Weaver[1] and Erwin Reifler[2] in early memoranda called attention to the problems of multiple meaning, while Oswald and Fletcher[3] began by fixing their attention on the word-order problems—particularly glaring in the case of German-to-English word-for-word translations. Over the years it has become increasingly clear that most, if not all, of the problems associated with word-for-word translation can be solved by the proper manipulation or utilization of the context. Context is to be understood here in its broadest interpretation. Contextual clues were treated in detail in an earlier article.[4] The six types of clues discussed there will be reformulated briefly here. They are:

1. *The field of discourse.* This was one of the earliest types of clues to be recognized. It can, by the use of specialized dictionaries, assist in the selection of the proper meaning of words that carry different meanings in different fields of discourse. The field of discourse may be determined by the operator, who places the appropriate glossary in the machine; or it may be determined by a machine routine on the basis of the occurrences of certain text words that are diagnostic of the field.

2. *Recognition of coherent word groups, such as idioms and compound nouns.* This clue can provide a basis for translating such word groups correctly even when their meaning does not follow simply from the meanings of the separate words.

3. *The syntactic function of each word.* If the translating program can determine syntactic function, clues will be available for solving word order problems as well as a large number of difficult multiple-meaning problems. Clues of this type will help, for example, in determining whether *der* in German should be translated as an article or as a relative or demonstrative pronoun, and whether it is nominative, genitive, or dative. They will also assist in handling the very difficult problems of translating prepositions correctly.

4. *The selectional relations between words in open classes, that is, nouns, verbs, adjectives, and adverbs.* These relations can be utilized by assigning the words to various meaning categories in such a way that when two or more of these words occur in certain syntactic relationships in the text, the correct meanings can be selected.

5. *Antecedents.* The ability of the translating program to determine antecedents will not only make possible the correct translation of pronouns, but will also materially assist in the translation of nouns and other words that refer to things previously mentioned.

6. *All other contextual clues, especially those concerned with an exact knowledge of the subject under discussion.* These will undoubtedly remain the last to be mechanized. Finding out how to use these clues to provide correct and accurate translations by machine presents perhaps the most formidable task that language scholars have ever faced.

Two Approaches

Attempts to learn how to utilize the above-mentioned clues have followed two separate approaches. One will be called the "95 percent approach" because it attempts to find a number of relatively simple rules of thumb, each of which will translate a word or

class of words correctly about 95 percent of the time, even though these rules are not based on a complete understanding of the problem. This approach is used by those who are seeking a short-cut to useful, if not completely adequate, translations. The other approach concentrates on trying to obtain a complete understanding of each portion of the problem so that completely adequate routines can be developed.

At any stage in the development of mechanical translation there will be some things that are perfectly understood and can therefore serve as the basis for perfect translation. In the area of verb, noun, and adjective inflection, it is possible to do a "100 percent job" because all the paradigms are available and all of the exceptions are known and have been listed. In this area one need not be satisfied with anything less than a perfect job. At the same time there will be some things about language and translation that are not understood. It is in this area that the difference between the two approaches shows up. The question of when to translate the various German, French, or Russian verb categories into the different sets of English verb categories is imperfectly understood. Those who adopt the 95 percent approach will seek simple partial solutions that are right a substantial portion of the time. They gain the opportunity of showing early test results on a computer. Those who adopt the 100 percent approach realize that in the end satisfactory mechanical translation can follow only from the systematic enlarging of the area in which we have essentially perfect understanding. The MIT group has traditionally concentrated on moving segments of the problem out of the area where only the 95 percent approach is possible into the area where a 100 percent approach can be used. Looking at mechanical translation in this light poses the greater intellectual challenge, and we believe that it is here that the most significant advances can be made.

Syntactic Translation

Examination of the six types of clues mentioned above reveals that they are predominantly concerned with the relationships of one word to another in patterns. The third type—the ability of the program to determine the syntactic function of each word—is basic to the others. It is basic to the first: If the machine is to determine correctly the field of discourse at every point in the text, even when the field changes within one sentence, it must use the relationship of the words in syntactic patterns as the key for finding which words refer to which field. It is basic to the second because idioms, noun compounds, and so on, are merely special patterns of words that stand out from more regular patterns. It is basic to the fourth because here we are dealing with selectional relationships between words that are syntactically related. It is basic to the fifth because the relationship of a word to its antecedent is essentially a syntactic relationship. It is probably even basic to the last, the category of all other contextual clues. Any approach to mechanical translation that attempts to go beyond mere word-for-word translation can with some justification be called a syntactic approach. The word "syntactic" can be used, however, to cover a number of different approaches. Following an early suggestion by Warren Weaver,[1] some of these take into consideration only the two or three immediately preceding and following words. Some of them, following a suggestion by Bar-Hillel,[5] do consider larger context, but by a complicated scanning forth and back in the sentence, looking for particular words or particular diacritics that have been attached to words in the first dictionary look-up. To the extent that these approaches operate without an accurate knowledge and use of the syntactic patterns of the languages, they are following the 95 percent approach. Oswald and Fletcher[3] saw clearly that a solution to the word-order problems in German-to-English translation required the identification of syntactic units in the sentence, such as nominal blocks and verbal blocks. Recently, Brandwood[6] has extended and elaborated the rules of Oswald and Fletcher. Reifler,[7] too, has placed emphasis on form classes and the relationship of words one with the other. These last three attempts seem to come closer to the 100 percent way of looking at things. Bar-Hillel,[8] at MIT, introduced a 100 percent approach years ago when he attempted to adapt to mechanical translation certain ideas of the Polish logician Ajdukiewicz. The algebraic notation adopted for syntactic categories, however, was not elaborate enough to express the relations of natural languages. Later, the author[9,10] proposed a syntactic method for solving multiple meaning and word-order problems. This routine analyzed and translated the input sentences in terms of successively included clauses, phrases, and so forth. More recently, Moloshnaya[11] has done some excellent work on English syntax, and Zarechnak[12] and Pyne[13] have been exploring with Russian a suggestion by Harris[14] that the text be broken down by transformations into kernel sentences which would be separately translated and then transformed back into full sentences. Lehmann,[15] too, has recently emphasized that translation of the German noun phrase into English will require a full descriptive analysis. In much of the work there has

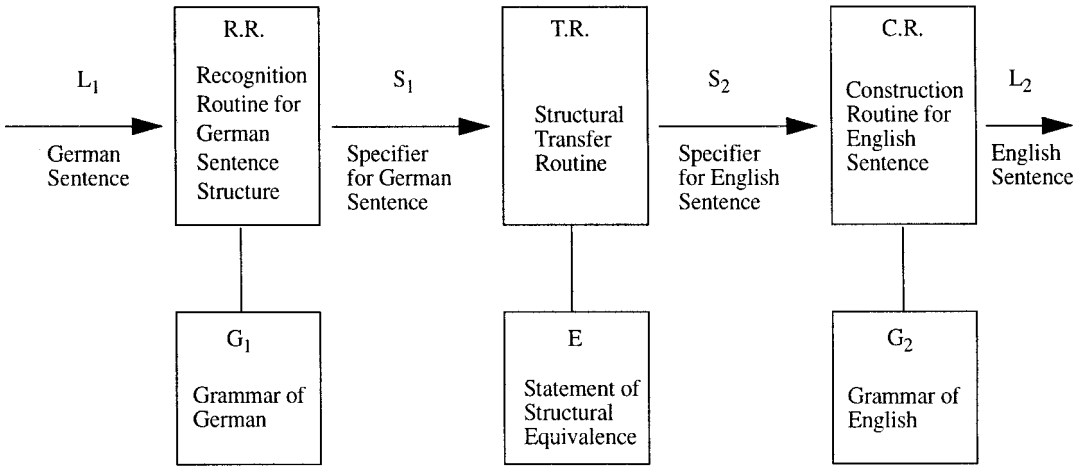


Figure 5.1
A framework for mechanical translation.

been an explicit or implicit restriction to syntactic relationships that are contained entirely within a clause or sentence, although it is usually recognized that structural features, to a significant extent, cross sentence boundaries. In what follows, we will speak of the sentence without implying this restriction.

The Framework

The framework within which we are working is presented in schematic form in figure 5.1. This framework has evolved after careful consideration of a number of factors. Foremost among these is the necessity of breaking down a problem as complex as that of mechanical translation into a number of problems each of which is small enough to be handled by one person.

Figure 5.1 represents a hypothetical translating machine. German sentences are fed in at the left. The recognition routine, R.R., by referring to the grammar of German, G_1, analyzes the German sentence and determines its structural description or specifier, S_1, which contains all of the information that is in the input sentence. The part of the information that is implicit in the sentence (tense, voice, and so forth) is made explicit in S_1. Since a German sentence and its English translation generally do not have identical structural descriptions, we need a statement of the equivalences, E, between English and German structures, and a structure transfer routine, T.R., which consults E and transfers S_1 into S_2, the structural description, or specifier, of the English sentence. The construction routine, C.R., is the routine that takes S_2 and constructs the appropriate English sentence in conformity with the grammar of English, G_2.

This framework is similar to the one previously published[16] except that now we have added the center boxes and have a much better understanding of what was called the "message" or transition language—here, the specifiers. Andreyev[17] has also recently pointed out that translation is essentially a three-step process and that current published proposals have combined the first two steps into one. One might add that some of the published proposals even try to combine all three steps into one. The question of whether there are more than three steps will be taken up later.

A few simple considerations will make clear why it is necessary to describe the structure of each language separately. First, consider the regularities and irregularities of declensions and conjugations. These are, of course, entirely relative to one language.

Context, too, is by nature contained entirely within the framework of one language. In considering the translation of a certain German verb form into English, it is necessary to understand the German verb form as part of a complex of features of German structure including possibly other verb forms within the clause, certain adverbs, the structure of neighboring clauses, and the like. In translating into English, the appropriate complex of features relative to English structure must be provided so that each verb form is understood correctly as a part of that English complex.

The form of an English pronoun depends on its English antecedent, while the form of a German pro-

noun depends on its German antecedent—not always the same word because of the multiple-meaning situation. As important as it is to locate the antecedent of the input pronoun in the input text, it is equally important to embed the output pronoun in a proper context in the output language so that its antecedent is clear to the reader.

In all of these examples it is necessary to understand the complete system in order to program a machine to recognize the complex of features and to translate as well as a human translator. If one is not able to fathom the complete system, one has to fall back on hit-or-miss alternative methods—the 95 percent approach. In order to achieve the advantages of full syntactic translation, we will have to do much more very careful and detailed linguistic investigation.

Stored Knowledge

The diagram (fig. 5.1) makes a distinction between the stored knowledge (the lower boxes) and the routines (the upper boxes). This distinction represents a point of view which may be academic: in an actual translating program the routine boxes and the stored knowledge boxes might be indistinguishable. For our purpose, however, the lower boxes represent our knowledge of the language and are intended not to include any details of the programming or, more particularly, any details of how the information about the languages is used by the machine. In other words, these boxes represent in an abstract fashion our understanding of the structures of the languages and of the translation equivalences. In an actual translating machine, the contents of these boxes will have to be expressed in some appropriate manner, and this might very well take the form of a program written in a pseudo code, programmable on a general-purpose computer. Earlier estimates[9] that the amount of storage necessary for syntactic information may be of the same order of magnitude as the amount of storage required for a dictionary have not been revised.

Construction

The Construction Routine, C.R. in figure 5.1, constructs to order an English sentence on the prescription of the specifier, S_2. It does this by consulting its pharmacopoeia, the grammar of English, G_2, which tells it how to mix the ingredients to obtain a correct and grammatical English sentence, the one prescribed.

The construction routine is a computer program that operates as a code conversion device, converting the code for the sentence, the specifier, into the English spelling of the sentence. The grammar may be looked upon in this light as a code book, or, more properly, as an algorithm for code conversion. Alternately, the construction routine can be regarded as a function generator. The independent variable is the specifier, and the calculated function is the output sentence. Under these circumstances, the grammar, G_2, represents our knowledge of how to calculate the function.

The sentence construction routine resembles to some extent the very suggestive sentence generation concept of Chomsky,[18] but there is an important difference. Where sentence generation is concerned with a compact representation of the sentences of a language, sentence construction is concerned with constructing, to order, specified sentences one at a time. This difference in purpose necessitates far-reaching differences in the form of the grammars.

Specifiers

For an input to the sentence construction routine, we postulated an encoding of the information in the form of what we called a specifier. The specifier of a sentence represents that sentence as a series of choices within the limited range of choices prescribed by the grammar of the language. These choices are in the nature of values for the natural coordinates of the sentence in that language. For example: to specify an English sentence, one may have to specify for the finite verb first, second, or third person, singular or plural, present or past, whether the sentence is negative or affirmative, whether the subject is modified by a relative clause, and which one, and so on. The specifier also specifies the class to which the verb belongs, and ultimately, which verb of that class is to be used, and so on, through all of the details that are necessary to direct the construction routine to construct the particular sentence that satisfies the specifications laid down by the author of the original input sentence.

The natural coordinates of a language are not given to us a priori, they have to be discovered by linguistic research.

Ambiguity within a language can be looked at as unspecified coordinates. A writer generally can be as unambiguous as he pleases—or as ambiguous. He can be less ambiguous merely by expanding on his

thoughts, thus specifying the values of more coordinates. But there is a natural limit to how ambiguous he can be without circumlocutions. Ambiguity is a property of the particular language he is using in the sense that in each language certain types of ambiguity are not allowed in certain situations. In Chinese, one can be ambiguous about the tense of verbs, but in English this is not allowed: one must regularly specify present or past for verbs. On the other hand, one is usually ambiguous about the tense of adjectives in English, but in Japanese this is not allowed.

It may be worth while to distinguish between structural coordinates in the narrow sense and structural coordinates in a broader, perhaps extralinguistic sense, that is, coordinates which might be called logical or meaning coordinates. As examples, one can cite certain English verb categories: in a narrow sense, the auxiliary verb "can" has two forms, present and past. This verb, however, cannot be made future or perfect as most other verbs can. One does not say "He has can come," but says, instead, "He has been able to come," which is structurally very different. It is a form of the verb "to be" followed by an adjective which takes the infinitive with "to." Again the auxiliary "must" has no past tense and again one uses a circumlocution—"had to." If we want to indicate the connection in meaning (paralleling a similarity in distribution) between "can" and "is able to" and between "must" and "has to," we have to use coordinates that are not structural in the narrow sense. As another example, there is the use of the present tense in English for past time (in narratives), for future time ("He is coming soon"), and with other meanings. Other examples, some bordering on stylistics, can also be cited to help establish the existence of at least two kinds of sentence coordinates in a language, necessitating at least two types of specifiers.

A translation routine that takes into consideration two types of specifiers for each language would constitute a five-step translation procedure. The incoming sentence would be analyzed in terms of a narrow structural specifier. This specifier would be converted into a more convenient and perhaps more meaningful broad specifier, which would then be converted into a broad specifier in the other language, then would follow the steps of conversion to a narrow specifier and to an output sentence.

Recognition

One needs to know what there is to be recognized before one can recognize it. Many people, including the author, have worked on recognition routines. Unfortunately, none of the work has been done with the necessary full and explicit knowledge of the linguistic structures and of the natural coordinates.

The question of how we understand a sentence is a valid one for linguists, and it may have an answer different from the answer to the question of how we produce a sentence. But it appears that the description of a language is more easily couched in terms of synthesis of sentences than in terms of analysis of sentences. The reason is clear. A description in terms of synthesis is straightforward and unambiguous. It is a one-to-one mapping of specifiers into sentences. But a description in terms of analysis runs into all of the ambiguities of language that are caused by the chance overlapping of different patterns: a given sentence may be understandable in terms of two or more different specifiers. Descriptions in terms of analysis will probably not be available until after we have the more easily obtained descriptions in terms of synthesis.

The details of the recognition routine will depend on the details of the structural description of the input language. Once this is available, the recognition routine itself should be quite straightforward. The method suggested earlier by the author[9] required that words be classified into word classes, phrases into phrase classes, and so on, on the basis of an adequate descriptive analysis. It operated by looking up word-class sequences, phrase-class sequences, and so on in a dictionary of allowed sequences.

Transfer of Structure

Different languages have different sets of natural coordinates. Thus the center boxes (fig. 5.1) are needed to convert the specifiers for the sentences of the input language into the specifiers for the equivalent sentences in the output language. The real compromises in translation reside in these center boxes. It is here that the difficult and perhaps often impossible matching of sentences in different languages is undertaken. But the problems associated with the center box are not peculiar to mechanical translation. Human translators also face the very same problems when they attempt to translate. The only difference is that at present the human translators are able to cope satisfactorily with the problem.

We have presented a framework within which work can proceed that will eventually culminate in mechanical routines for full syntactic translation. There are many aspects of the problem that are not yet understood and many details remain to be worked out.

We need detailed information concerning the natural coordinates of the languages. In order to transfer German specifiers into English specifiers, we must know something about these specifiers. Some very interesting comparative linguistic problems will undoubtedly turn up in this area.

The author wishes to express his indebtedness to his colleagues G. H. Matthews, Joseph Applegate, and Noam Chomsky for some of the ideas expressed in this paper.

Notes

This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Research and Development Command), and the U.S. Navy (Office of Naval Research); and in part by the National Science Fouocation.

l. Warren Weaver, "Translation," *Machine Translation of Languages*, ed. William N. Locke and A. Donald Booth, Wiley, New York and London (1955).

2. Erwin Reifler, "Studies in Mechanical Translation No. 1, MT," mimeographed (January 1950).

3. Victor A. Oswald, Jr., and Stuart L. Fletcher, Jr., "Proposals for the Mechanical Resolution of German Syntax Patterns," *Modern Language Forum*, Vol. XXXVI, No. 2–4 (1951).

4. V. H. Yngve, "Terminology in the Light of Research on Mechanical Translation," *Babel*, Vol. 2, No. 3 (October 1956).

5. Y. Bar-Hillel, "The Present State of Research on Mechanical Translation," *American Documentation*, Vol. 2, pp. 229–237 (1951).

6. A. D. Booth, L. Brandwood, and J. P. Cleave, *Mechanical Resolution of Linguistic Problems*, Academic Press, New York (1958).

7. Erwin Reifler, "The Mechanical Determination of Meaning," *Machine Translation of Languages*, ed. William N. Locke and A. Donald Booth, Wiley, New York and London, 1955.

8. Y. Bar-Hillel, "A Quasi-Arithmetical Notation for Syntactic Description," *Language*, Vol. 29, No. 1 (1953).

9. V. H. Yngve, "Syntax and the Problem of Multiple Meaning," *Machine Translation of Languages*, ed. William N. Locke and A. Donald Booth, Wiley, New York and London, 1955.

10. V. H. Yngve, "The Technical Feasibility of Translating Languages by Machine," *Electrical Engineering*, Vol. 75, No. 11 (1956).

11. T. N. Moloshnaya, "Certain Questions of Syntax in Connection with Machine Translation from English to Russian," *Voprosy Yazykoznaniya*, No. 4 (1957).

12. M. M. Zarechnak, "Types of Russian Sentences," *Report of the Eighth Annual Round Table Meeting on Linguistics and Language Studies*, Georgetown University (1957).

13. J. A. Pyne, "Some Ideas on Inter-Structural Syntax," *Report of the Eighth Annual Round Table Meeting on Linguistics and Language Studies*. Georgetown University (1957).

14. Z. S. Harris, "Transfer Grammar," *International Journal of American Linguistics*, Vol. XX, No. 4 (October 1954).

15. W. P. Lehmann, "Structure of Noun Phrases in German," *Report of the Eighth Annual Round Table Meeting on Linguistics and Language Studies*. Georgetown University (1957).

16. V. H. Yngve, "Sentence-for-Sentence Translation," *Mechanical Translation*, Vol. 2, No. 2 (1955).

17. N. D. Andreyev, "Machine Translation and the Problem of an Intermediary Language," *Voprosy Yazykoznaniya*, No. 5 (1957).

18. Noam Chomsky, *Syntactic Structures*, Mouton, The Hague (1957).

6
The Present Status of Automatic Translation of Languages

Yehoshua Bar-Hillel

1 Aims and Methods, Survey and Critique

[. . .]

1.2 Unreasonableness of Aiming at Fully Automatic High Quality Translation

During the first years of the research in MT, a considerable amount of progress was made which sufficed to convince many people, who originally were highly skeptical, that MT was not just a wild idea. It did more than that. It created among many of the workers actively engaged in this field the strong feeling that a working system is just around the corner. Though it is understandable that such an illusion should have been formed at the time, it was an illusion. It was created, among other causes, also by the fact that a large number of problems were rather readily solved, and that the output of machine-simulated "translations" of various texts from Russian, German or French into English were often of a form which an intelligent and expert reader could make good sense and use of. It was not sufficiently realized that the gap between such an output, for which only with difficulty the term "translation" could be used at all, and high quality translation proper, i.e., a translation of the quality produced by an experienced human translator, was still enormous, and that the problems solved until then were indeed many but just the simplest ones, whereas the "few" remaining problems were the harder ones—very hard indeed.

Many groups engaged in MT research still regard fully automatic, high quality translation (FAHQT) as an aim towards which it is reasonable to work. Claims to the effect that FAHQT from Russian to English is attainable in the near future were recently made, for instance, by one of the four subgroups working on MT at Georgetown University (section 2.1.3). I shall discuss these claims below. But let me state already at this point that I could not be persuaded of their validity. On the contrary, I am quite ready to commit myself to concoct Russian sentences or, should this for some reason be regarded as unfair, to exhibit actually printed Russian sentences for which a perusal of the proposed translation program of this group, or of any other group that would offer in the near future a method of fully automatic translation, would result either in gibberish or, what is even worse, in meaningful but wrong translations. I am so convinced of this because I believe to be in possession of an argument which amounts to an almost full-fledged demonstration of the unattainability of FAHQT, not only in the near future but altogether. This demonstration is given in appendix III.

Most groups, however, seem to have realized, sometimes very reluctantly, that FAHQT will not be attained in the near future. Two consequences can be drawn from this realization. One can go on working with FAHQT in mind, in the hope that the pursuit of this aim will yield interesting theoretical insights which will justify this endeavor, whether or not these insights will ever be exploited for some practical purpose. Or one gives up the ideal of FAHQT in favor of some less ambitious aim with a better chance of attainability in the near future. Both consequences are equally reasonable but should lead to rather different approaches. Lack of clarity in this respect, vague hopes that somehow or other both aims can be attained simultaneously and by the use of the same methods, must lead to confusion and result in waste of effort, time and money. Those who are interested in MT as a primarily practical device must realize that full automation of the translation process is incompatible with high quality. There are two possible directions in which a compromise could be struck; one could sacrifice quality or one could reduce the self sufficiency of the machine output. There are very many situations where less than high quality machine output is satisfactory. There is no need to present examples. If, however, high quality is mandatory—and I do not think, for instance, that scientists are prepared to be satisfied with less than the present average standard of human translation, while many regard this standard as too low for their purposes—then the machine output will have to be post-edited,

thereby turning, strictly speaking, machine translation into *machine aids to translation.*

1.3 Commercial Partly Mechanized, High Quality Translation Attainable in the Near Future

In the remainder of this survey, I shall exclusively deal with those situations where translation involved has to be of high quality. It should be easy to see how the conclusions at which I arrive have to be modified in order to deal with situations in which lesser quality is satisfactory.

As soon as the aim of MT is lowered to that of high quality translation by a *machine–post-editor partnership*, the decisive problem becomes to determine the region of optimality in the continuum of possible divisions of labor. It is clear that the exact position of this region will be a function of, among other things, the state of linguistic analysis to which the languages involved have been submitted. It may be safely assumed that, with machine-time/efficiency becoming cheaper and human time becoming more expensive, continuous efforts will be made to push this region in the direction of reducing the human element. However, there is no good reason to assume that this region can be pushed to the end of the line, certainly not in the near future.

It seems that with the state of linguistic analysis achieved today, and with the kind of electronic computers already in existence or under construction, especially with the kind of large capacity, low cost and low access-time internal memory devices that will be available within a few years, a point has been reached where commercial partly mechanized translation centers stand a serious chance of becoming a practical reality. However, various developments are still pending and certain decisions will have to be made.

First, a reliable and versatile mechanical print reader will have to become available. It has been estimated that the cost of retyping printed Russian material into a form and on a medium that could be processed by a machine would amount, under present conditions, to about one fourth of a cent per word [7]. This estimate is probably too low, as the quality of the retyping has to be exceptionally high, in order to avoid printing mistakes which would perhaps be quite harmless for a human reader, but could be rather disastrous for machines which so far are totally unable to deal with misprints. The original text might therefore have to be keypunched by two operators, verified, etc., or else to be keypunched once, but at highly reduced speed. Indeed, whereas the above estimate is based on a rate of 20 Russian words per minute, another report [8] gives the maximum rate of trained and experienced keypunch operators as half this number. In one place [9], it is estimated that an automatic print reader might be ten times cheaper than human retyping. The difference between one half of a cent per keypunched word and one twentieth of a cent per print-read word could make all the difference, as the present cost per word of human Russian-to-English translation in the United States is generally given as lying between one and three cents [10], apparently depending on the quality and urgency of the job, and perhaps also on the exact form of the output. The costs may be different, of course, for other language pairs and in other countries. An informative synopsis on the variation of rates of payment for scientific and technical translation is given in a recent UNESCO survey [11].

Secondly, a concerted effort will have to be made by a pretty large group in order to prepare the necessary dictionary or dictionaries in the most suitable form. That this is not such a straightforward affair as laymen are apt to think becomes clear in the work of the Harvard MT group [12, 13]. This group developed an interesting semiautomatic method for preparing dictionaries (section 2.1.6).

Thirdly, a good amount of thinking accompanied by an equally large amount of experimenting will have to go into the determination of the location of the interval in the above-mentioned continuum within which the optimal point of the division of labor between machine and post-editor will have a good chance of being situated, as a function of the specific translation program and the specific qualities of the envisaged post-editor. Among other things, these studies would have to determine whether some minimal pre-editing, while requiring but very little knowledge of the source language by the pre-editor, could not be utilized in order to reduce the load of the machine by a considerable amount. At present, many of the experimental MT programs make use of such limited pre-editing (section 2.1.4). As one illustration of an operation that is in almost all cases so ridiculously simple for a human pre-editor that it could be almost instantaneously performed by a keypunch operator with only the barest knowledge of the source language, let me mention the distinction between the functioning of a point as a period, hence as one of the all-important markers of end-of-sentence, and its various other functions. Having the machine make

this decision—a vital one, indeed so vital that it is one of the first operations, if not the first, in many translation programs that shun the use of pre-editing altogether—might be a complex and costly affair, throwing some doubts on the soundness of the case presented above in favor of a mechanical print reader. For the time being, at least, so long as keypunching is being used for the input, it is doubtless profitable to introduce as much elementary pre-editing as the keypunch operator can take into stride without considerably slowing down.

Fourthly, an old question which has not been treated so far with sufficient incisiveness, mostly because the ideal of FAHQT diverted the interests of the research workers into other, less practical directions, namely the question whether MT dictionaries should contain as their source language entries all letter sequences that may occur between spaces, sometimes called *inflected forms*, or rather so-called *canonical forms*, or perhaps something in between like *canonical stems* [14], has to be decided one way or other before mass production of translations is taken up. This question is clearly highly dependent, among other things, upon the exact type of internal and external memory devices available, and it is therefore mandatory to have a reliable estimate of this dependence. It is obvious that the speed of the machine part of the translation, and thereby the cost of the total translation process, will depend to a high degree on the organization of the dictionaries used. Most workers in the field of MT seem to have rather definite, though divergent, opinions in this respect. However, I am not aware of any serious comparative studies, though the outcome of such studies most surely will have a considerable impact upon the economics of MT.

In general, the intention of reducing the post-editor's part has absorbed so much of the time and energy of most workers in MT, that there has not been sufficient discussion of the problem whether partially automatic translation, even with such a large amount of participation by the post-editor as would be required under present conditions, is not nevertheless a desirable and feasible achievement. I fully understand the feeling that such an achievement is not of very high intellectual caliber, that the real challenge has thereby not yet been taken up, but I do not think that those agencies for whom any reduction of the load imposed at the moment on the time of highly qualified expert translators is an important achievement, should necessarily wait with the installation of commercial man-machine translation outfits until the post-editor's part has become very small, whatever amount of satisfaction the MT research worker will get from such an achievement. It is gratifying to learn that this attitude coincides with that of the Harvard group (section 2.1.6) and is probably now shared by many other groups in the USA, USSR, and England, though it would further the issue if clear-cut statements of policy could be obtained in this respect.

1.4 Compromising in the Wrong Direction

At this stage, it is probably proper to warn against a certain tendency which has been quite conspicuous in the approach of many MT groups. These groups, realizing that FAHQT is not really attainable in the near future so that a less ambitious aim is definitely indicated, had a tendency to compromise in the wrong direction for reasons which, though understandable, must nevertheless be combated and rejected. Their reasoning was something like the following: since we cannot have 100% automatic high quality translation, let us be satisfied with a machine output which is complete and unique, i.e., a smooth text of the kind you will get from a human translator (though perhaps not quite as polished and idiomatic), but which has a less than 100% chance of being correct. I shall use the expression "95%" for this purpose since it has become a kind of slogan in the trade, with the understanding that it should by no means be taken literally. Such an approach would be implemented by one of the two following procedures: the one procedure would require to print the most frequent target-language counterpart of a given source-language word whose ambiguity has not been resolved by the application of the syntactical and semantical routines, necessitating, among other things, large scale statistical studies of the frequency of usage of the various target renderings of many, if not most, source-language words; the other would be ready to work with syntactical and semantical rules of analysis with a degree of validity of no more than 95%, so long as this degree is sufficient to insure uniqueness and smoothness of the translation. This approach seems wrong to me and even dangerous since the machine output of the corresponding program will be of low quality in a misleading and soothing disguise. Since so many sentences, "5%" of a given text, will have a good chance of being mistranslated by the machine, it is by no means clear whether the reader will always be able to detect these mistranslations, just because the machine output is so smooth and

grammatical (so let us assume for the sake of the argument, though I doubt whether even this much can really be at this stage of the game) that he might be able to find only few cues to warn him that something is wrong with it. It is not inconceivable that the machine translation would be so wrong at times as to lead its user to actions which he would not have taken when presented by a correct translation. (When I talk about "100%" I obviously have in mind not some heavenly ideal of perfection, but the product of an average qualified translator. I am aware that such a translator will on occasion make mistakes and that even machines of a general low quality output will avoid some of these mistakes. I am naturally comparing averages only.)

But there is really no need at all to compromise in the direction of reducing the reliability of the machine output. True enough, a smooth machine translation looks impressive, especially if the reader is unable to realize at first sight that this translation is faulty ever so often, but this esthetically appealing feature should not blind us to see the dangers inherent in this approach. It is much safer to compromise in the other direction. Let us be satisfied with a machine output which will every so often be neither unique nor smooth, which every so often will present the post-editor with a multiplicity of renderings among which he will have to take his choice, or with a text which, if it is unique, will not be grammatical. On the other hand, whenever the machine output is grammatical and unique it should be, to adopt a slogan current in the Harvard group, "fail-safe" (to about the same degree, to make this qualification for the last time, as the average qualified human translator's output is fail-safe). Let the machine by all means provide the post-editor with all possible help, present him with as many possible renderings as he can digest without becoming confused by the *embarras de richesse*—and here again we have quite a problem of finding an interval of optimality—but never let the machine make decisions by itself on purely frequential reasons even if these frequencies can be relied upon. If these frequency counts could be done cheaply—and I doubt very much whether this is feasible to such a high degree of reliability as would probably be required for our purposes—let this information too be given the post-editor, but by no means should practical MT wait until this information is obtained.

The only reasonable aim, then, for short-range research into MT seems to be that of finding some machine–post-editor partnership that would be commercially competitive with existing human translation, and then to try to improve the commercial effectiveness of this partnership by improving the programming in order to delegate to the machine more and more operations in the total translation process which it can perform more effectively than the human post-editor. These improvements will, of course, utilize not only developments in hardware, programming (especially automatic programming), and linguistic analysis, but also the experience gained by analyzing the machine output itself. Should it turn out that for the sake of competitiveness some use of a pre-editor, and perhaps even of a bilingual post-editor, would be at least temporarily required, then this fact should be accepted as such, in spite of the trivialization of the theoretical challenge of the MT problem which would be entailed by such a procedure.

1.5 A Critique of the Overestimation of Statistics and the "Empirical Approach"

Let me finish this part of the survey by warning in general against overestimating the impact of statistical information on the problem of MT and related questions. I believe that this overestimation is a remnant of the time, seven or eight years ago, when many people thought that the statistical theory of communication would solve many, if not all, of the problems of communication. Though it is often possible by a proper organization of the research effort to get a certain amount of statistical information at no great extra cost, it is my impression that much valuable time of MT workers has been spent on trying to obtain statistical information whose impact on MT is by no means evident. It is not true that every statistic on linguistic matters is automatically of importance for MT so that the gathering of any such statistics could be regarded as an integral part of MT research without any need for additional justification.

Gathering of statistics is regarded by many MT groups as being part of a more general methodological approach—the so-called "*empirical approach*" [15]. This term has already caused a lot of confusion. I am using it here in the sense in which it is employed by the RAND group [16]. This sense should become obvious from the following discussion. Adherents of this approach are distrustful of existing grammar books and dictionaries, and regard it as necessary to establish from scratch the grammatical rules by which the source-language text will be machine analyzed, through a human analysis of a large enough corpus of source-language material, constantly improving upon the formulation of these rules by constantly

enlarging this corpus. With regard to dictionaries, a similar approach is often implemented and a dictionary compiled from translations performed by bilingual members of the group or by other human translators considered to be qualified by this group. This approach seems to me somewhat wasteful in practice and not sufficiently justified in theory. The underlying distrust seems to have been caused by the well-known fact that most existing grammars are of the normative type, hence often of no great help in the analysis of actual writing (and to an even higher degree, of actual speech), and that existing dictionaries are of such a nature that quite often none of the presented target-language counterparts of a source-language word are satisfactory within certain contexts, especially with regard to terms used in recently developed scientific fields. However, even in view of these facts, I believe that the baby has far too often been thrown away with the bathwater. No justification has been given for the implicit belief of the "empiricists" that a grammar satisfactory for MT purposes will be compiled any quicker or more reliably by starting from scratch and "deriving" the rules of grammar from an analysis of a large corpus than by starting from some authoritative grammar and changing it, if necessary, in accordance with analysis of actual texts. The same holds *mutatis mutandis* with regard to the compilation of dictionaries. But grammars have in general not wholly been dreamt up, nor have dictionaries been compiled by some random process. Existing grammars and dictionaries are already based, though admittedly not wholly, upon actual texts of incomparably larger extension than those that serve as a basis for the new compilers. Russian is not Kwakiutl, and with all due regard to the methods and techniques of structural linguistics and to the insights which this science has given us in respect to some deficiencies of traditional grammars, I do not think that it follows from its teachings that all existing codifications of languages with a highly developed literature should be totally disregarded. Let me add, without going here into details for lack of space, that the empiricalness of the derivations of grammar rules from actual texts is rather doubtful as such. For certain general methodological considerations one might as well be led to the conclusion that these rules incorporate a lot of subjective and highly biased and untested assumptions such that their degree of validity might very well, on the average, be lower than that of the well-established, often-tested and critically examined grammars, in spite of their normativity.

2 Critical Survey of the Achievements of the Particular MT Research Groups

After these far too short (and therefore occasionally rather dogmatic) general comments, it is now time for a more detailed survey of the approaches and achievements of the twenty or so groups which are at present actively engaged in research on MT or on linguistic topics believed to be of immediate relevance for MT. In one case a defunct group (section 2.1.5) is being mentioned, first because it made significant contributions during its existence, and secondly because there is still some chance that it may be revived. This survey will deal exclusively with the more general aspects of the MT problem and especially with research methodology. Therefore, the innumerable specific advance of the various groups with regard to coding, transliterating, keypunching, displaying of output, etc., will be mentioned only rarely. But the list of references should contain sufficient indications for the direction of the reader interested in these aspects.

The order in which these groups will be discussed is: USA, Great Britain, USSR, others, following, with one exception, the order of degree of my personal acquaintance. Within each subdivision, the order will in general be that of seniority.

2.1 The USA Groups

2.1.1 The Seattle Group Professor Erwin Reifler of the University of Washington, Seattle, started his investigations into MT in 1949, under the impact of the famous memorandum by Weaver [17], and has since been working almost continuously on MT problems. The group he created has been constantly increasing in size and is at present one of the largest in the States. In February 1959, it published a 600-page report describing in detail its total research effort. This report has not reached me at the time of writing this survey (April 1959) which is the more unfortunate as the latest publication stemming from this group is a talk presented by Reifler in August 1957 [18], and I was, due to a personal mishap, unable to visit Seattle during my stay in the States. It is not impossible that my present discussion is considerably behind the actual developments.

The efforts of this group seem to have concentrated during the last years on the preparation of a very large Russian-English automatic dictionary containing approximately 200,000 so-called "operational entries" whose Russian part is probably composed of what was termed above (section 1.3) "inflected forms" (as against the million or so inflected forms

corresponding to the total Russian vocabulary of one hundred thousand canonical forms). This dictionary was to be put on a photoscopic memory device, developed by Telemeter-Magnetics Inc. for the U.S. Air Force, which combines a very large storage capacity with very low access time and apparently is to be used in combination with one of the large electronic computers of the IBM 709 or UNIVAC 1105 types. The output of this system would then be one version of what is known as *word-by-word translation*, whose exact form would depend on the specific content of the operational entries and the translation program. Both are unknown to me though probably given in the above mentioned report. Word-by-word Russian-to-English translation of scientific texts, if pushed to its limits, is known to enable an English reader who knows the respective field to understand, in general, at least the gist of the original text, though of course with an effort that is considerably larger than that required for reading a regular high quality translation, or else to enable an expert English post-editor to produce on its basis, with some very restricted use of the original text (in transliteration, if he does not know how to read Cyrillic characters), a translation which is of the same order of quality as that produced by a qualified human translator. However, no comparisons as to quality and cost between the Seattle MT system and human translation are given in the publications known to me. In any case, in view of the rather low quality of the machine output (word-by-word translation is theoretically a triviality, of course, though a lot of ingenuity is required to get the last drop out of it), the claim that the Seattle–Air Force system is "the most advanced translation system under construction" [19] is very misleading; even more misleading is the name given the photoscopic disc, "The USAF Automatic Language Translator Mark I" [20], which creates the impression of a special purpose device, which it is not.

The Seattle group started work towards getting better-than-word-by-word machine output in the customary direction of automatically changing the word order and reducing syntactical and lexical ambiguities (the Seattle group prefers to use the terms "grammatical" and "non-grammatical") but again little is known of actual achievements. One noticeable exception is Reifler's treatment of German compound words, which is an especially grave problem for MT with German as the source-language since this way of forming new German nouns is highly creative so that the machine will almost by necessity have to identify and analyze such compounds [21]. In the above-mentioned 1957 talk, Reifler claimed to have "found moreover that only three matching procedures and four matching steps are necessary (sufficient?) to deal effectively with—that is, to machine translate correctly—any of these ten types of compounds of any[!] language in which they occur," [22]—a claim which sounds hardly believable, whose attempted substantiation is probably contained in the mentioned report. It is worthwhile to stress that this group does not adopt the "empirical approach" mentioned above, and is not going to be satisfied with so-called "representative samples," but is trying to keep in view the ascertainable totality of possible constructions of the source language, though representative samples are of course utilized during this process [23].

For reasons given above, I must strongly disagree with Reifler's "belief that it will not be very long before the remaining linguistic problems in machine translation will be solved for a number of important languages" [24]. How dangerous such prophecies are is illustrated by another prophecy of Reifler's, to the effect that "in about two years (from August 1957) we shall have a device which will at one glance read a whole page and feed what it has read into a tape recorder and thus remove all human cooperation on the input side of the translation machines" [25]. The best estimates I am aware of at present mention five years as the time after which we are likely to have a reliable and versatile print reader (section 1.3) at the present rate of research and development.

2.1.2 The MIT Group I started work on MT at the Research Laboratory of Electronics of MIT in May 1951. In July 1953, when I returned to Israel, Victor H. Yngve took over, steadily recruiting new assistants for his research. During the last years, the MIT group has laid great stress on its adherence to the ideal of FAHQT. For this purpose they regard the complete syntactical and semantical analysis of both source- and target-language to be a necessary prerequisite. It is, therefore, to these processes that their research effort has been mostly directed. It seems that this group is aware of the formidableness of its self-imposed task, and is rather uncertain in its belief that this prerequisite will be attained in the near future. In one of his latest publications, Yngve says: "It is the belief of some in the field of MT that it will eventually be possible to design routines for translating mechanically from one language into another without human intervention" [26]. It is rather obvious from the context that Yngve includes himself among the "some."

How remote "eventually" and "ultimately"—another qualifying adverb occurring in a similar context—are estimated to be is not indicated. On the other hand, the MIT group believes, and I think rightfully, that the insights into the workings of language obtained by its research are valuable as such, and could at least partly be utilized in practical lower aimed machine translation by whomever is interested in this latter aim. However, it will probably be admitted by this group that some of the research undertaken by it might not be of any direct use for practical MT at all. The group employs to a high degree the methods of structural linguistics, and is strongly influenced by the recent achievements of Professor Noam Chomsky in this field [27].

The impact upon MT of Chomsky's recently attained insights into the structure of language is not quite clear. Since I presented my own views on this issue in a talk at the Colloque de Logique, Louvain, September 1958 [28], as well as in a talk given before the Second International Congress of Cybernetics, Namur, September 1958, a greatly revised version of which is reproduced in appendix II, I shall mention here only one point. The MIT group believes, I think rightly, that Chomsky has succeeded in showing that the *phrase structure model* (certain variants of which are also known as *immediate constituent models*) which so far has served as the basic model with which structural linguists were working, in general as well as for MT purposes, and which, if adequate, would have allowed for a completely mechanical procedure for determining the syntactical structure of any sentence in any language for which a complete description in terms of this model could be provided—as I have shown for a weak variant of this model, already 6 years ago [29] by a method that was later improved by Lambek [30]—is not fully adequate and has to be supplemented by a so-called *transformational model*. This insight of Chomsky explains also, among other things, why most prior efforts at the mechanization of syntactical analysis could not possibly have been entirely successful. The MIT group now seems to believe that this insight can be given a positive twist and made to yield a more complex but still completely mechanical procedure for syntactical analysis. I myself am doubtful about this possibility, especially since the exact nature of the transformations required for an adequate description of the structure of English (or any other language) is at the moment still far from being satisfactorily determined. A great number of highly interesting but apparently also very difficult theoretical problems, connected with such highly sophisticated and rather recent theories as the theory of recursive functions, especially of primitive recursive functions, the theory of post-canonical systems, and the theory of automata (finite and Turing), are still waiting for their solution, and I doubt whether much can be said as to the exact impact of this new model on MT before at least some of these problems have been solved. I think that Chomsky himself cherishes similar doubts, and as a matter of fact my present evaluation derives directly from talks I had with him during my recent visit to the States.

The MIT group has, among other things, also developed a new program language called COMIT which, though specially adapted for MT purposes, is probably also of some more general importance [31], and whose use is envisaged also by other groups. The fact that it was felt by this group that a program language is another more or less necessary prerequisite for MT is again the result of their realization of the enormous difficulties standing in the way of FAHQT. It is doubtful whether the development of a program language beyond some elementary limits is indeed necessary, or even helpful for more restricted goals. I would, however, agree that a program language is indeed necessary for the high aims of the MIT group, though I personally am convinced that even this is not sufficient, and that this group, if it continues to adhere to FAHQT, will by necessity be led in the direction of studying learning machines. I do not believe that machines whose programs do not enable them to learn, in a sophisticated sense of this word, will ever be able to consistently produce high-quality translations.

About the actual achievements of the MIT group with regard to MT proper little is known, apparently due to its reluctance to publish incomplete results. It is often felt that because of this reluctance other MT workers are wasting some of their time in treading over ground that might have already been adequately covered, though perhaps with negative results.

2.1.3 The GU Group The largest group working on MT in the States is that at Georgetown University, Washington, D.C., led by Professor Dostert. The GU group comprises four subgroups. One of these is headed by Professor Garvin and has been engaged during the last two years exclusively in programming the mechanization of the syntactical analysis of Russian. Their method seems to work rather satisfactorily for the syntactical analysis of a large class of Russian sentences, though its exact reach has not yet been fully determined nor all the details of their program

debugged. They have produced a very large number of publications, in addition to a multitude of Seminar Work Papers of the Machine Translation Project of Georgetown University, of which I shall mention only two of the more recent ones [32, 33].

The other three subgroups at GU are working on MT as a whole, two of them from Russian into English, the third from French into English. It is claimed that during the last few months the research done at GU has broadened and MT from additional languages into English has begun to be investigated. However, I am not aware of any publications reporting on these new activities and shall therefore not deal with them here. They seem to be at present in their preliminary stages only.

I already mentioned above (section 1.2) that far-reaching claims were made by one of the GU subgroups. This is the group headed by Miss Ariadne W. Lukjanow and using the so-called *Code Matching Technique* for the translation of Russian chemical texts. I expressed then my conviction that this group could not possibly have developed a method that is as fully automatic and of high quality as claimed. There are in principle only two procedures by which such claims can be tested. The one consists in having a rather large body of varied material, chosen by some external agency from the field for which these claims are made, processed by the machine and carefully comparing its output with that of a qualified human translator. The other consists in having the whole program presented to the public. None of these procedures has been followed so far. During a recent demonstration mostly material which had been previously lexically abstracted and structurally programmed was translated. When a text, lexically abstracted but not structurally programmed, was given the machine for translation, the output was far from being of high quality and occasionally not even grammatical. True enough, this did not prevent the reader from understanding most of the time what was going on, but this would have been the case also for word-by-word translation, since the sample, perhaps due to its smallness, did not contain any of those constructions which would cause word-by-word translation to be very unsatisfactory. In contrast, however, with word-by-word translation which, if properly done, is hardly ever wrong, though mainly only because it is not real translation and leaves most of the responsibility to the post-editor, this translation contained one or two rather serious errors, as I was reliably told by someone who carefully went through the machine output and compared it with the Russian original. (I myself did not attend the demonstration, and my knowledge of Russian is rather restricted.)

The task of evaluating the claims and actual achievements of the Lukjanow subgroup is not made easier by the fact that there seems to exist only one semipublicly available document prepared by herself [34]. This document contains 13 pages and is not very revealing. The only peculiarity I could discover lies in the analysis of the source-text in a straight left-to-right fashion, in a single pass, exploiting each word as it comes, including the demands it makes on subsequent words or word blocks, whereas most other techniques of syntactical analysis I know go through the source-language sentences in many passes, usually trying to isolate certain units first. I shall return to Miss Lukjanow's approach below (section 2.1.9).

The claim for uniqueness (and adequacy) of the translation of a chemical text is based upon an elaborate classification of all Russian words that occurred in the analyzed corpus into some 300 so-called *semantical classes*. Though such a detailed classification should indeed be capable of reducing semantic ambiguity, I am convinced that no classification will reduce it to zero, as I show in appendix III, and that therefore the claim of the Lukjanow group is definitely false. There should be no difficulty for anyone who wishes to take the trouble to exhibit a Russian sentence occurring in a chemical text, which will be either not uniquely translated or else wrongly translated by the Lukjanow procedure, within a week after all the details of this procedure are in public possession.

On the other hand, I am quite ready to believe that this subgroup has been able to develop valid techniques for a *partial* mechanization of Russian-to-English high quality translation of chemical literature (or else for a full mechanization of low quality translation), but, unfortunately, this group seems to be extremely reluctant to make the details of its program publicly available. Should it turn out that they did make some real progress not achieved elsewhere, this reluctance will have caused a great waste of time and money in other MT research groups.

A third subgroup at GU, led by Dr. Michael Zarechnak, proceeding in a somewhat different manner, using a so-called *General Analysis Technique*, and is making less far-reaching claims. Much of its work which I was able to check seemed to me well-founded and to contain solid achievements. However, as this is not the place to go into technical details, it is not possible to present an exact evaluation of where this subgroup stands right now. The only fully official

publication of this group [35]—the number of semi-official Seminar Work Papers is rather large—is too short to tell the whole story. Dr. Zarechnak hopes to be ready with a demonstration within a few months, and I also understand that everybody is welcome to look over the program to the degree that it has already been written up. The group does envisage the utilization of a post-editor for high quality final output.

With regard to the fourth and last subgroup at GU, led by Dr. A. F. R. Brown, I shall say very little here since I was unable to talk with Brown personally. As already mentioned, he is mostly interested in translation from French into English. I understand from his numerous Seminar Work Papers that he is developing his program on a sentence-after-sentence basis, i.e., dealing with the translation problems as they come and, so I was told, solving them one after another with great ingenuity. I have already expressed my conviction that this approach is somewhat wasteful and am sorry indeed that I was unable to talk the issue over with one of the seemingly most successful adherents of the empirical approach.

Altogether, I think that among themselves the four subgroups at GU cover a good deal of the problems arising in connection with MT. Dostert's interest in this field stems from his participation in the first MT Conference in June 1952, and so does Garvin's who attended the public opening meeting of that conference. These two linguists have been spending much of their time since on scientific and organizational aspects of MT, and have succeeded in training a large number of other people now working on MT at GU. This is a good deal of experience, and it is therefore not surprising that the work done under their direction should indeed cover most aspects of the MT problem. I am stressing this point since, in spite of the fact that I do disagree with some of the views and approaches of Dostert and his collaborators, I believe that every newcomer to the field—and there have been many of those during the last year and more are in prospect—should make himself as thoroughly acquainted as possible with the work done at GU, and get as clear a picture as possible of their achievements and failures. Otherwise he will have a good chance of repeating work that has been done there, and perhaps repeating the many failures that undoubtedly must have occurred there during the years. There exists no other group in the United States, or in England for that matter, which has been working on such a broad front. This remark of mine is not to be interpreted as implying that the prospective newcomers will not have to get acquainted with anything done outside GU. On the contrary, I do not think that there is much done at GU in the field of MT which is not being done also elsewhere, sometimes in more than one place, and in some of these places perhaps even more effectively. But GU is still a good place to get a full view of the problem or rather could be so if each subgroup were equally willing to discuss in full detail its work with others.

2.1.4 The RAND Group The RAND Corporation in Santa Monica, California, became interested in MT in 1949, and has dealt with MT off and on since. The well known study by Professor Abraham Kaplan on the reduction of ambiguity through context [36] was done at RAND, and Dr. Olaf Helmer of RAND participated in the first MT Conference. However, it is only during the last years that RAND's interest in MT has greatly increased so that the present RAND MT group headed by Dr. David G. Hays, with Professor Kenneth E. Harper of UCLA serving as its chief consultant, is at the moment one of the larger ones. It is there that the empirical approach has found its perhaps strongest expression, probably because Harper is such a strong believer in its soundness. The method they advocate is to go over a certain sample of Russian texts, say of 30,000 words in length, "derive" from a human analysis of this corpus both a dictionary and a set of syntactical and semantical rules, test the derived dictionary and rules on a new sample of the same size, increase the dictionary and, if necessary, expand as well as improve upon the rules as a result of this test, go on to the next sample, etc. As a matter of fact, during the first six cycles—by now, they might have finished the eighth—they have mostly tried to perfect the dictionary and solve some of the problems of polysemy (this is the term preferred by the Russian authors; it is certainly more convenient than the terms "semantical ambiguity," "lexical ambiguity," "non-grammatical ambiguity," "multiple meanings" used in Anglo-Saxon countries), for example, that bothersome problem of the unique rendering of Russian prepositions. It is only a few months ago that they started to attack the question of syntactical analysis. It is impossible to here go into a detailed description of their planned approach but, again, it is quite empirical and therefore rather slow, and not too promising in its details as they stand at the moment. So, for instance, it is planned to investigate hundreds of thousands of Russian consecutive word pairs in order to arrive at a revealing classification of such pairs for the purpose of reducing syntac-

tical ambiguity, to be followed by an investigation of word triplets, etc. This procedure is, of course, rather natural and consciously, or unconsciously, based upon the immediate constituent model discussed above. I am not too much impressed by the claim that resolution of syntactical ambiguities by consideration of the immediate neighborhoods of the ambiguous expression has proved itself in practice. Not that I doubt that syntactical ambiguity as well as polysemy, for that matter, can quite often be completely resolved and even more often be considerably reduced through the exploitation of the immediate neighborhood of the ambiguous expression; it is only that I have what I regard to be good theoretical reasons for believing that this reduction will in general stop short of complete resolution. I am here talking of only those cases where the original sentence is not syntactically ambiguous as such; if it is ambiguous, then a technique that would "resolve" this ambiguity rather than display it would be no good. These theoretical reasons are, for syntactical ambiguity, that the immediate constituent model is not fully adequate so that trying to push a resolution technique based upon it beyond certain limits (which at present are admittedly by no means clear) must defeat itself. In addition, and this is probably of greater practical importance, I cannot persuade myself that "deriving" syntactical rules from a huge number of observations will yield better and quicker results than testing rules, whatever their "derivation," as to their ability to stand up against concocted counter-examples. This seems to me to be a methodological commonplace. In physics or chemistry, it is now probably generally agreed that it is a much better methodology to put "freely conceived" theoretical constructions to as sharp empirical tests as possible in order to refute these constructions than to arrive at theories that are only compatible with existing observations. I see no reason why linguistics should be different in this respect and why in this field observation and "derivation" is to be regarded as superior to empirical testing of "freely conceived" theories. A report which is probably inspired by Harper, if not actually written by him, contains a statement to the effect that its author is not very much impressed by the fact that counter-examples to his empirically derived rules can be concocted so long as these are concocted examples and not ones that occur in some actual text [37]. Final judgment on this issue must be left to the reader.

Most results of the RAND research are being published in a series of nine research memoranda called "Studies in Machine Translation," six of which had appeared between December 1957 and October 1958, with a possibility that the remaining three might have appeared in the meantime. I have already had an opportunity to mention one of these studies above (section 1.1) and to praise its general reliability[1][6].

Another study contains a very clear statement of the RAND research methodology [38], the remainder being various manuals dealing with such matters as instructions for transliteration, coding, keypunching, pre-editing and post-editing of Russian scientific texts, all of them of great practical interest but outside the scope of this survey, as stated above (section 1.5). Special mention is deserved, however, of the memorandum [39] containing a list of 225 Russian articles in Physics and Mathematics, comprising 227,752 running words, that are available at RAND in punched cards for the use of system and procedure designers who might require textual material for their research. Part of the keypunching was done by the Ann Arbor group (section 2.1.7).

2.1.5 The Ramo–Wooldridge Group In the area of greater Los Angeles there was another group working on MT at Ramo–Wooldridge Corporation. It started operating in 1955 and was directed in its last stage by Dr. Don R. Swanson. Harper acted as a consultant for this group at an earlier stage, and in 1957–58 there existed a close cooperation between the RAND group and the Ramo–Wooldridge one. Though there are some differences between their approaches, a description of these differences would require going into greater detail than I am prepared to do here. The Ramo–Wooldridge group published two highly interesting reports [7, 10], to both of which I have already had an opportunity to refer. A close study of these reports should be of great help to everybody in the field. I understand that work on MT at Ramo–Wooldridge has been discontinued at the end of 1958, though perhaps only temporarily so.

2.1.6 The Harvard Group The Harvard University group, headed by Professor Anthony G. Oettinger, stands in many respects quite apart from the others. First, it has busied itself for years almost exclusively with an exploration of the word-by-word translation method. Secondly, this preoccupation was accompanied by, and originated partly out of, a strong distrust of the achievements of other groups. Though it must be admitted that the possibilities of word-by-word translation from Russian into English have never before been so thoroughly explored as they were by this group, with many new insights gained, and that very valuable results were obtained as to

the structure and construction of MT dictionaries, one may still wonder whether this group really struck the golden middle between utilizing other people's work in the field and distrusting their work, though there certainly were good reasons for the distrust on quite a few occasions.

The progress made by this group can be easily evaluated by comparing two doctoral theses submitted at Harvard University, the one—to my knowledge the first dissertation on MT—by Oettinger [40] in 1954, the other by Giuliano in January 1959 [41]. This second thesis seems to close an era and indicate the opening of a new one. The first five chapters describe the operation of the Harvard Automatic Dictionary, the methods for its compiling and updating, as well as a great variety of applications, in such thoroughness and detail that the impression is created that not much more is to be said on this subject. The last chapter, on the other hand, contains some interesting but tentative and almost untested remarks on what Giuliano calls a *Trial Translator* [42], i.e., an automatic programming system for the experimental production of better than word-by-word translations.

Out of the enormous amount of material contained in this thesis, let me dwell on those passages that are of immediate relevance to the question of the commercial feasibility of MT. The existing program at the Harvard Computation Laboratory can produce word-by-word Russian-to-English translations at a sustained rate of about 17 words per minute on a UNIVAC I, and about 25 words per minute on a UNIVAC II. This is 4–6 times more than an expert human translator can produce, but since UNIVAC II time is 100 times more expensive than a human translator's time, commercial MT is out of the question at present. Giuliano estimates that a combination of an IBM 709 (or UNIVAC 1105) with the photoscopic disc mentioned above (section 2.1.1) would, after complete reprogramming—requiring some three programmer years—and a good amount of other development work, be able to produce translations at 20–40 times the present rate which, taking into account the increase in the cost of computer time, would still leave the cost of a word-by-word machine translation slightly above that of a high-quality human translation. The difference will, however, now be so slight that one may expect that any further improvement, in hardware and/or in programming, would reverse the cost relationship. This does not yet mean that true word-by-word MT will be in business. The cost of post-editing the word-by-word output in order to turn it into a passable translation of the ordinary type would probably be not much less than producing a translation of this quality without machine aid. As a matter of fact, senior research scientists having excellent command of scientific Russian and English, and extensive experience in technical writing, would be hampered rather than assisted by the automatic dictionary outputs in their present form. The number of these individuals is, on the other hand, rather small and few of them can take the time from their scientific work to do a significant amount of translating and would have to be remunerated several times the ordinary professional translator's fee to be induced to spend more time on translating.

Altogether, it does not seem very likely that a non-subsidized, commercial translation service will, in the next five years or so, find use for an automatic dictionary as its only mechanical device. However, as the Harvard group is quick to point out, an automatic dictionary is an extremely valuable research tool with a large number of possible applications, some of which have already proved their value. Let me add that in situations where speed is at a premium, high quality is not a necessary requisite, and human translators at a shortage for any price such situations might arise, for instance, in military operations—automatic dictionaries would be useful as such for straight translation purposes.

The whole issue is, however, somewhat academic. There is no need to speculate what the commercial value of an automatic dictionary would be since the same computer-store combination that would put out a word-by-word translation can be programmed to put out better than word-by-word translations. This is, of course, the subject on which most MT groups, including the Harvard group itself as of this year, are working on right now. At what stage a winning machine-post-editor combination will be obtained, is not so easy to foresee. I personally believe that the combination of a computer of twice the efficiency/cost ratio of an IBM 709 and Photographic Disk Store, of a program resulting from the pooled present knowledge of all the groups working on Russian-to-English MT, and of 2–3 years concentrated effort on improving the input and output should turn the trick, but I admit that so many factors are involved which have not yet been adequately evaluated that this belief is somewhat irrational.

2.1.7 The Ann Arbor Group There is probably no topic among those studied by the MT group working in the Willow Run Laboratories of the University of Michigan, Ann Arbor, and directed by Mr. Andreas

Koutsoudas, which is not also studied elsewhere. This is by no means meant to be derogatory; the same remark applies to almost every other MT group as well. The particular achievements could still be unique.

The group adopted, in general, a methodology similar to the one prevalent at RAND (section 2.1.4)—there has been close collaboration between these groups—though with much more attention to theoretical models [43]. In addition to keypunching a large corpus of Russian scientific articles with frequency counts, members of the group are working on a resolution of syntactical and lexical ambiguity by context, as well as on algebraic and automata theory models of language, following the lead of Chomsky (section 2.1.2).

2.1.8 The Philadelphia Group The group at the University of Pennsylvania, Philadelphia, headed by Professor Zellig S. Harris, is wholly concerned with developing programs for the syntactical analysis of English, without bothering at every step with the implications of its research for MT. They do, however, intimate that their research will lead to useful applications for MT, and even more so for information retrieval [44] and related problems. These intimations are based upon serious misinterpretations of the semantic impact of their own work and are to that degree unsubstantiated. Though the actual programs compiled by the Philadelphia group for the syntactic analysis of English embody solid achievements based upon valid intuitive insights as well as upon extremely painstaking and detailed observations, and are in this respect equal if not superior to parallel achievements obtained during the same period by other groups concerned with the same problem (or rather, in most cases, with the materially different but methodologically very similar problem of mechanically analyzing the structure of Russian, German, French, etc.), the theory behind these achievements seems to be of doubtful validity, if interpreted literally, and ill-formulated and misleading in any case. No detailed substantiation of this rather harsh judgment can, of course, be undertaken here. A few comments must do.

Harris introduced into linguistic theory the terms "transformation" and "kernel." These terms are, unfortunately, not at all well-defined and rely for their meaning on a farfetched and underdeveloped analogy with the use of these terms in modern abstract algebra. (It is worthwhile stressing here, in view of current misunderstandings, that Harris' use has to be carefully distinguished from that of his former pupil Chomsky with whom these terms have well-determined and clear meanings and are free of any pseudomathematical flavor. Chomsky is, of course, highly influenced by the views of Harris and vice versa—but his formulations are certainly exempt from the shortcomings criticized here.) It is, however, obvious from the context that Harris regards a sentence and its negation, or a sentence and its passive, as mutual transforms of each other. So, for example, *Atoms emit electrons* and *Atoms do not emit electrons* are transforms of each other, as are *Atoms emit electrons* and *Electrons are emitted by atoms*, with *Atoms emit electrons* being the kernel in both cases. It is therefore surprising to read that "a sentence carries the same information as does its transform" and that "a sentence, or a text, transformed into a sequence of kernels carries approximately the same information as did the original" or even that "for scientific, factual, and original material, however, it seems that the relevant information is held constant under transformation, or is varied in a way that depends explicitly on the transformation used" [45]. As against these formulations disregarding their internal inconsistency—it must be pointed out that not only do negations of kernel sentences not carry, in general, the same, or even approximately the same, information as the kernel sentences themselves, but that it is not even true that the information carried by kernel sentences is varied by negation, or by the passive transformation, in the same way. Though *Atoms emit electrons* and *Electrons are emitted by atoms* carry the same information, *Some number exceeds every number* and *Every number is exceeded by some number* do not. (As a matter of fact, the first sentence is false, and the second true.) Harris was apparently partly aware of this, since he later [46] treats *not* both as part of the kernel, rather confusingly and inconsistently with his definition—though due to its vagueness this cannot be stated very definitely—and as an operator on the kernel. (The question of applicability to MT of the analysis of a given sentence as the outcome of the performance of zero, one or more transformations, in a certain order, on one or more kernel sentences—these terms now taken in Chomsky's sense—I have dealt with in appendix II.)

In order not to be misunderstood, let me stress that my criticism refers only to Harris' description of what the process he calls *kernelization* is apt to achieve and that part of his theory of transformations which lies behind it. From a short discussion with him, I gathered that some of his formulations are indeed not to be understood literally, but I was unable to determine

what exactly was left. It would be of some importance to get more clarity on this issue.

On the other hand, the actual programming of the mechanization of the syntactical analysis of English, as produced by the Philadelphia group, is utterly independent of the transformational model and, so far at least, based exclusively on the immediate constituent model. Judging from its latest internal project reports, amazingly great progress has been made. Similar to the Harvard group, however, the programming is hampered by the fact that the internal low-access-time memory of the UNIVAC I, with which it works, is too small for MT purposes.

2.1.9 The NBS Group Coming now to the four "young" groups that started their activities within the last two years, let me first briefly describe the work done by the group consisting of Mrs. Ida Rhodes and a couple of associates at the National Bureau of Standards, Washington, D.C. No publications exist so far. It is nevertheless my opinion, based upon a few talks during which I was able to go through her program in considerable detail, that her approach is promising and worth close study. Not that she has been able so far to achieve any new results, but she quite often reobtained old results by sufficiently new and occasionally quite ingenious methods. Mrs. Rhodes is one of the few people in the field who has had long experience with actual programming. Being a native Russian speaker, she succeeded in combining her linguistic intuitions with her thorough knowledge of computers and their programming into an MT program which, judging from its presently existing outline, should, when fully developed, be able to achieve much of what can be achieved in this field in one of the most efficient and economical ways of which I am aware. Mrs. Rhodes is a mathematician by training, and her knowledge of modern structural linguistics is very slight. It should furnish some grounds for thought to realize how much of the practical aims of MT can be attained with so little use of structural linguistics. It should, however, be taken into account that Mrs. Rhodes' aims are wholly practical, and that no attempt is made by her to obtain a FAHQ output.

Let me mention just one detail in her program. One of the major problems in the syntactical analysis of the given source-language sentence is the problem of where to start. Garvin, for example, instructs the machine to look first for participial constructions and relative clauses. Harris, working with English though, lets the machine look for nominal blocks beginning with the end of the sentence and working backwards. In both approaches it is necessary to go over the sentence a few times before its final analysis is obtained; as a matter of accident, three passes are envisaged by both Garvin and Harris. Mrs. Rhodes, perhaps because of her linguistic naïveté, starts the analysis always with the first word of the sentence and lets the machine go over the words one after another, each time rewriting part of its own program, recalling Miss Lukjanow's technique mentioned above (section 2.1.3). Though Mrs. Rhodes' approach is then based, as it were, on a *finite-state model* for the detailed description of which the reader is referred to Chomsky's booklet [47]—whereas Garvin and Harris are working in effect with a phrase structure model, which is demonstrably a more powerful one, it is interesting to note that this does not interfere, apparently, with its practical efficiency. As against the multipass technique of Garvin and Harris, it has the advantage of being much more easily transferable to the treatment of the translation from other languages whereas, I presume, Harris' and Garvin's approaches are very much more tailored to English and Russian, respectively. In this connection, the interesting question arises, which of these three procedures is closest to the one used by human translators, if human translators use one common procedure at all which seems to me to be at least highly doubtful. Not that this question is of any practical importance for MT at this moment; however, if and when the time comes when translations will be performed by machines with learning abilities and using, at least partly, rather general heuristic instructions instead of the fully spelled-out program which is customary at present, our question may become a practical one since we would then probably want to give the machine the same or similar heuristic instructions which are given today to human translators during their training or which they develop for themselves in time. [. . .]

2.2 The British Groups

2.2.1 The London Group One of the two British MT groups is operating at Birkbeck College in London and is directed by Dr. Andrew D. Booth. Booth is one of the very first persons who thought of the utilization of electronic computers for translation as early as 1946. In 1948 he wrote, together with Dr. R. H. Richens, presently a member of the Cambridge group, a pioneer paper on MT [51]. He has continued his research in this field almost uninterruptedly though always only part-time, and published last

year, together with two associates, a book dealing mostly with MT [52]. He was also coeditor of the first book dealing with machine translation [53], which contained 14 monographic studies on various aspects of the MT problem, in addition to a foreword by Dr. Warren Weaver of the Rockefeller Foundation and a valuable historical introduction. His recent book contains a great wealth of insights into the syntactical structure of German, and to a lesser degree into that of French and Russian, but the approach suffers from an excessive adherence to the empirical method in so much as rules for resolving syntactical ambiguity are based, in principle, "on analysis of all the existing literature on the subject in question," and in practice, for the purposes of illustration, on the analysis of a very small amount of text. The same holds for the methods proposed in this book for the reduction of semantical ambiguities. The authors are aware of the limitations of this method but intend to leave the development of a method that would resolve ambiguities in all conceivable (scientific) texts to people with a high degree of acquaintance with the German language. Some of the statements made in this book, of either historical or systematic nature, are made in an offhand manner and could create a somewhat distorted picture, especially with regard to the relative importance of the insights gained by the London group itself. There is, however, no point here of going into such details. The book contains, in addition, many technical details on the construction of programs for MT, a full account of which may be gained from a companion volume by Mrs. Booth [54].

It might be worth mentioning that this book also contains a refutation of one very frequent argument favoring the use of an artificial mediating language, an *interlingua* in short, for MT purposes [55]. This argument points out that translation from each of n natural languages into each other requires the establishment of $n(n-1)$ programs (including dictionaries), whereas the use of an interlingua, into which and from which all translation exclusively proceeds, requires only $2n$ such programs. (For ten languages, for example, this means a reduction from 90 to 20 programs.) The fallaciousness of this argument is immediately obvious, however, as soon as one realizes that using one, any one, of the original n languages as a mediating language would reduce the number of programs even more, namely to $2(n-1)$ (in our illustration to 18). This counter-argument does not, of course, prove that the idea of using an interlingua for MT purposes is wrong as such, since other arguments might be brought forward in its support, but the one refuted just now seems to have been one of the most potent ones, and with its elimination proponents of the interlingua idea should give it a second thought.

It should indeed be carefully tested, for independent reasons, to what degree the quality of a translation between two languages is impaired, if instead of a direct translation an indirect one is employed, based upon high-quality translation from the source-language into some intermediate language and from it into the target-language. So far there exist, to my knowledge, only more or less anecdotal results in this respect. Should it turn out that high quality translation is generally obtainable by going through some intermediate language, natural or artificial, this would be of enormous importance for multilingual MT of the future.

Whereas the mentioned argument "from n^2 to $2n$" for the use of an *artificial interlingua* in MT can definitely be proven fallacious, though it holds good as an argument for the use of *any* intermediate language, there are of course other arguments to support the use of an artificial interlingua *qua* artificial whether of the Esperanto type or of that of a symbolic language system. I admit that the idea of a "logical," unambiguous (in every respect, morphologically, syntactically, and semantically) interlingua has its appeal today as had the related idea of a *characteristica universalis* in the seventeenth and early eighteenth centuries. This appeal is bolstered by the great achievements of modern mathematical logic with its constant use of artificial language systems, and there is therefore some force in the claim that an idea that failed in the 17th century need not do so in the 20th. But the present argument is no less fallacious. Its fallacy lies in the assumption that "translation" from a natural language into a "logical" one is somehow simpler than translation from one natural language into another. This assumption, however, is totally unwarranted, whatever its appeal to someone with little direct experience with symbolic language systems. As a matter of fact, the transition from a sentence in a natural language to its counterpart in a language system deserves the name "translation" only in a somewhat Pickwickian sense. I shall not elaborate this point any further, but only mention that it has been discussed rather widely in recent methodological literature. The fallacy is probably another result of the customary loose use of the word "translation" which has already caused a lot of trouble on other occasions (such as in connection with information retrieval where the issue becomes constantly befuddled through an uncritical and still more meta-

phorical use of this word). Not only is the process of presenting a counterpart of some natural language sentence in some symbolic language system in general more difficult than its translation into some other natural language, even for a human being, as everyone who has ever taught a freshman course in symbolic logic will readily certify, but the mechanization of this kind of "translation" poses problems which are much more difficult than those posed by translation proper. It is no accident, again, that not only have linguists not attacked these problems in any serious sense, but that even hard-boiled logicians have shunned it in favor of dealing with "easier" ones (which ordinary linguists regard as lying beyond their comprehension). Altogether, the problems revolving around an interlingua as a device for MT are still in a highly speculative state, and it is probable that years will pass before any practical results can be expected.

2.2.2 The Cambridge Language Research Unit The second British group is located in Cambridge, England, and is directed by Miss Margaret Masterman. [...] In spite of its constant disclaimers, this group is a highly speculative one with many of the good and equally many of the less good connotations of the term. I find myself again and again amazed by the prolificy of ideas emerging from it, almost all of which have some initial appeal, while also having the disturbing property of constantly changing their exact meaning or being quickly replaced by some other idea, for which the same process starts all over again after a very short time. I myself, in the early stages of my thinking on MT, played with many of these ideas and can therefore readily testify to their appeal. I did, for instance, spend some time on the question of whether and to what degree *Combinatorial Logic* [56] could be applied to MT, and though I have failed so far to achieve any results in this connection, I am not convinced that I myself, or other people better equipped for this purpose, could not still do so if working very hard and uninterruptedly on this problem. In one of my publications I made a brief mention of this issue [57]. Miss Masterman wrote a long (unpublished) paper on this topic three years ago, but I had great trouble understanding its point, and the issue is no longer mentioned in more recent publications of the Cambridge group, having apparently been superseded by the idea of applying *Lattice Theory* [58, 59]. Now Lattice Theory is the theory of a structure which is so general that one should not be surprised to find it embodied in many actual situations. There can also be no doubt that lattice theory, and certain still more general branches of abstract algebra such as the theory of semilattices, trees, directed graphs and partially ordered systems, can be applied to linguistic investigations though I am not aware of any new insights gained so far by such applications. The applications made by the Cambridge group of their lattice-theoretical approach, inasmuch as they are valid, are only reformulations in a different symbolism of things that were said and done many times before.

A third idea emerging from this group, though not only from it, is that of using a *Thesaurus*-type dictionary in lieu of, or perhaps in addition to, ordinary dictionaries. I find here the greatest difficulties of understanding in spite of many attempts on my part to do so and many hours of talking with various members of the group. One cause of the troubles is the fact that the term "thesaurus" has not only been used by various groups in different, occasionally quite different, senses, but that members of the same group often use the term in different senses, and that its meaning keeps shifting even in the publications of one and the same person with no adequate warning given to the reader, perhaps without the writer being aware of such a shift. So we find that a thesaurus is sometimes meant to be rather similar to Roget's well-known *Thesaurus of the English Language*, and at other times to be rather different from it. Sometimes the thesaurus is supposed to contain after each entry so many expressions of the same language, at other times its equivalents in some interlingua [60–62]. Since I could not persuade myself that I really understood the Cambridge group's conception (or conceptions?) of the thesaurus (or thesaurus-lattice) approach to MT, I shall say nothing about it. Perhaps the reader will be luckier. The literature cited above and the further references contained there should suffice for this purpose.

By far the most important idea cherished by the Cambridge group is that of using an interlingua for MT purposes. In addition to what was already said above on this topic the following remarks might help to explain the attraction that this idea has exercised in the Cambridge group as well as in many of the Russian groups (though not, to any noticeable degree, in any of the American groups). It is undoubtedly true that human translators occasionally (and beginners quite often) use a procedure that might be described as consisting of three separate steps: coming across the German word *schreibt*, for instance, a translator sometimes explicitly (and may be said to always proceed this way implicitly, so long as this way of speaking is not pushed beyond certain limits) first analyzes this word as third person, singular, present,

active of *schreiben*, then looks up *schreiben* in the German-English dictionary, finding *write* as its English rendering, and finally synthesizes the third person, singular, present, active of *write* as *writes*. When attempting to render the same word in French, he may perform three completely analogous steps: he first analyzes *schreibt* as third person, singular, present, active of *schreiben*, then looks up *schreiben* in the German-French dictionary, finding *écrire* as its French rendering and finally synthesizes the third person, singular, present, active of *écrire* as *écrit*. If some book has to be machine translated into English and French at the same time, it seems, therefore, that it would be more economical to perform the respective first steps in common. Generalizing, one might come to the conclusion that it would be less economical to perform every machine translation in these three steps of *analysis*, *one-to-one transfer*, and *synthesis*, than to perform the first step which is purely a source-language affair independently of the target-language into which the text is going to be translated, and the third step independently of the source-language from which the text was translated. Generalizing still further, one might think of proceeding this way not only with regard to the treatment of single words (morphological analysis and synthesis) but also with regard to whole phrases and sentences (syntactical analysis and synthesis).

There can be no doubt as to the appeal of this idea, and I am quite sure that every MT worker must have thought of it sooner or later, probably sooner as I did, for instance, in the first week of my preoccupation with this topic. Would it not be an enormous saving indeed to have, when dealing with multiple mutual translation between n languages, instead of the required $n(n-1)$ translation programs and $n(n-1)$ unidirectional binary dictionaries—analysis programs, n synthesis programs and *one* completely symmetrical n-ary dictionary? Hence, why waste this time on preparing one-way translation programs for one pair of languages at a time, having to start from scratch for each new pair of languages and even for the same pair translating in the opposite direction?

There is only one thing wrong with this idea, which relies, of course, on another, more complex variant of the argument "from n^2 to $2n$"; it is utterly chimerical and based upon a series of fallacies, and the only effect of its adoption would be that of postponing for an indefinite time, depending on n, the date of inaugurating the first commercial partly mechanized translation center. First, the idea of a completely symmetrical n-ary dictionary, each entry consisting of exactly n words, one each for each of the n languages concerned, is wholly unrealistic. The Cambridge group is fully aware of this fact, but in its attempt to find a substitute, it has been led to the idea of an interlingual thesaurus of undetermined complexity. On the other hand, to the degree that a symmetrical n-ary idioglossary is practical—as it might conceivably be for some highly restricted technical field—its preparation would require exactly the same effort as the preparation of the $n-1$ (two-way) binary idioglossaries from one of the n languages to all the others.

The situation is similar with regard to the analysis and synthesis part. The morphological and syntactical categories in terms of which an analysis of a given source-language is successful for translation into some given target-language might not be the best ones for some other target-language. Analyzing in terms of all possible categories that might be needed for translation into any of the $n-1$ other languages is exactly of the same degree of complexity as the sum of the analyses for each language in turn. The illusion that this is not so is created by regarding the effort of preparing a "complete," "absolute" grammatical analysis of one language as being only slightly more complex than preparing such an analysis "relative" to some other language, and simultaneously regarding the preparation of $n-1$ "relative" grammatical analyses as requiring only slightly less than $n-1$ times the effort required for preparing one such analysis. The truth, of course, is that the effort is on the average exactly the same, almost by definition.

That the terms "interlingua," "intermediate language," "mediating language"—and their counterparts in Russian—are being used in many different senses should not come as any special surprise. Natural languages, artificial languages of the Esperanto type, symbolic language-systems of the type treated by logicians, "algebraic" languages of various denominations, all have been suggested at one time or other as candidates for mediating languages. I believe that my present criticisms hold equally against each of these interpretations. I find myself here in close agreement with the views of Booth [62a].

As already said above (section 2.2.1), the only point in this connection that deserves serious consideration is the following: Assuming that translation programs already exist from language L_1 to L_2 and from L_2 to L_3, by how much would the output of a direct translation program from L_1 to L_3 be better than the output of a combination of the two existing programs, and would the difference pay for the effort required to

prepare the new program? It seems to me almost trivially true that to this question no general theoretical answer is possible. It just depends.

Altogether there exists so far no evidence that any of the ideas brought forward by the various members of the Cambridge group will ever contribute new effective methods for practical MT, and little evidence that they would result in new valid insights into the workings of language. [. . .]

2.3 The USSR Groups

2.3.1 General In Soviet Russia there are some ten groups active in MT research. In contradistinction to the United States, however, where each group has access to an electronic computer, some of the Russian groups have so far apparently not been able to put their theoretical schemes to an actual machine test. Whether this is because of a shortage of available machine time or of a relatively larger interest in theoretical analysis, I do not know. It may be due to both. Experimental testing of theories is quite often performed by having human beings simulate machines, as was often done in the States six or seven years ago.

In addition to more or less permanent MT groups there are many scientists who apparently do not belong to any of these groups, but spend much of their time on MT research, sometimes serving as consultants, occasionally to many groups simultaneously and taking an active part in the rather frequent conferences and Academy of Sciences meetings dedicated wholly or partly to MT.

Active research on MT in Russia started in January 1955 on a relatively large scale from the very beginning. It is likely that the interest of the government in MT was stirred up by the GU demonstration in January 1954. However, MT had a prehistory in Russia. It seems that the first serious attempt to mechanize translation was made in 1933 by an engineer, P. P. (Smirnov-) Troyansky who proposed in that year to construct "a machine for selecting and printing words by translation from one language into another or into several others simultaneously," and even got an author's certificate for his invention [63]. His proposal met with scepticism and derision on behalf of the Russian linguists and mathematicians of the time and fell into oblivion. It is hard to blame them in view of the fact that the advent of electronic digital computers was still a dozen years ahead. Troyansky seems to have died at the end of World War II. One may express the hope that more should become known of this Babbage of MT.

In general, it is rather safe to assume that the state of MT in Russia is not essentially different from that elsewhere, and that the direction which MT research is taking there is about the same as, say, in the United States, in spite of some statements to the contrary made by Russian MT workers in 1956 and later. If anything, Russian scientists may be somewhat ahead in the linguistic analysis, whereas they are probably somewhat behind in actual machine-testing. We find in Russia the same differences in policy as elsewhere, between the adherents of FAHQT and those who advocate more modest aims, between the theoretically-minded and the empirically-minded, etc.

Though all this is true in general, there are two points in which MT research in Russia seems to be different from that in the States by degree, though not qualitatively. First, and this point has gotten much publicity, probably more than it objectively deserves, the number of language pairs between which MT has been investigated is much larger in the USSR research. Morphological and syntactical analyses for MT, automatic dictionaries, and comparative studies of linguistic structure for MT have been carried out with varying depth for some twenty languages, and MT programs exist for perhaps thirty pairs though most of these programs are still rather rudimentary, with Russian always, and quite naturally, being either the target or the source language. Among these languages we find, as one would expect, the great world languages of science, English, French, and German, languages of countries with whom the USSR has political alliances, such as Chinese, Czech, Bulgarian, and Albanian, but also quite a number of non–Indo-European languages (in addition to Chinese) with totally different scripts, such as Japanese, Hindi, Arabic, Indonesian, Vietnamese, etc. It is somewhat surprising to see that MT from Norwegian, for instance, has also already been investigated. This is hardly to be explained by the existence of an urgent practical need in this direction. It is more likely that MT is often used as a pretext for the instigation of descriptive studies of languages from a structuralist point of view—the only one which makes sense for MT—a view which until a few years ago was rejected in Russia as "formalistic." In this respect, again, the situation is not much different from that prevailing in the USA where a good amount of solid linguistic research is carried out under the auspices, not to say disguise, of MT. I understand that all this is by no means typical for "applied" linguistics alone.

If we recall that in the USA and England, English is always the target language (with the exception of

the GU group where English-to-Chinese and English-to-Japanese MT is in the first stages of investigation) and that the only source languages considered so far at any large scale were Russian, French and German exclusively, the impression on the general public created by the breadth of Russian MT is understandable. More important, however, is that Russian has been treated as both target and source language. I am not referring now to the political and cultural aspects of this fact, but rather to the fact that this state seems to have been one of the strongest inducements for the second issue which distinguishes Russian MT, though again only in degree.

This second issue is the almost universal preoccupation with intermediate languages. The use of this term in Russia is in general no less equivocal and confusing than in England. One Russian group, the most practical and down-to-earth of them (section 2.3.2), would like to investigate multiple translation with Russian as the pivot language. I already expressed above (sections 2.2.1, 2.2.2) my belief that this aspect of the intermediate language issue is indeed one, probably the only one, that deserves serious consideration.

Other groups, however, are using the term "intermediate language" in different senses, some of which are closely related to—though conceived independently of—those in which it is understood by the Cambridge group. Though considerably less fantastic in the details than the Cambridge speculations, there is nothing in the Russian versions of the intermediate language idea to change my sceptical opinion as to its practical value. [...]

Before I go into a relatively detailed discussion of the achievements of the particular MT groups, let me report briefly on some general organizational aspects of MT. I already mentioned in the first paragraph of this survey (section 1.1) that in the May 1958 First All-Union MT Conference in Moscow, 79 institutions were represented. These included 21 institutes of the Academy of Sciences of the USSR, 8 institutes of the Academies of Sciences of the Union Republics, 11 universities and 19 other institutions of higher learning in the country. This does not mean, of course, as it would not for similar conferences in the USA, that there exist active research groups in all these institutions. Still, this conference was undoubtedly the largest of its kind held so far. Not less than 71 talks were given during its week-long meeting, though not all of them were of direct concern to MT.

A seminar on mathematical linguistics has been in operation since September 1956 in the Department of Philology of the Moscow State University. An Association for Machine Translation was established in December 1956 at the First Moscow State Pedagogical Institute of Foreign Languages and is publishing a *Bulletin of the Seminar on Problems of Machine Translation* (*Byulleten ob''edineniya po problemam mashinnogo perevoda*). A Committee on Applied Linguistics was established in June 1958, with its center apparently in Leningrad. It was decided to establish permanent liaison between the Committee and the Association. The Russian biweekly, *Voprosy Yazykoznaniya* (*Problems of Linguistics*), carried in the years 1956–58 a large number of papers dedicated to MT and continues to deal with MT as a regular feature. In 1958, *Problemy Kibernetiki* (*Problems of Cybernetics*) started to appear, with A. A. Lyapunov as editor, and the first issue contained three papers on MT.

The best over-all description of Soviet work in MT was given in a talk presented by V. Yu. Rozentsveyg at the Fourth International Congress of Slavicists, Moscow, 1958 [68]. The second edition of a booklet by D. Yu. Panov on automatic translation [69] goes into much greater details than Rozentsveyg's talk but concerns itself, on the other hand, mostly with the achievements of the ITMVT group. [...]

[A survey of contemporary MT groups in USSR is omitted, as the material is described in Oettinger's contribution in this volume—*Eds.*]

3 Conclusion

Fully automatic, high quality translation is not a reasonable goal, not even for scientific texts. A human translator, in order to arrive at his high quality output, is often obliged to make intelligent use of extralinguistic knowledge which sometimes has to be of considerable breadth and depth. Without this knowledge he would often be in no position to resolve semantical ambiguities. At present no way of constructing machines with such a knowledge is known, nor of writing programs which will ensure intelligent use of this knowledge.

Reasonable goals are then either fully automatic, low quality translation or partly automatic, high quality translation. Both are theoretically feasible and, for certain language pairs, attainable today though not yet on a commercial scale. Through a concentration of effort, pooling of knowledge, and planned division of labor it should be possible to establish within a period of three to five years translation centers which would be in a position, after a

short period of subsidized operation, to make considerable use of electronic machinery in the translation process under competitive costs and with a substantial saving in expert bilingual manpower. It might perhaps be possible, after some additional time, to get along without expert bilingual post-editors altogether. For high quality output, however, the services of a human monolingual editor—in general, for practical reasons, a post-editor, though I can see no reasons why a pre-editor or even a pair of post- and pre-editors could not be equally effective, wherever available, and therefore deplore the almost total disregard of these possibilities—will remain indispensable. I regard it as unlikely that the cost of human editing can be pushed down in the near future below half of the cost of human translating of equal quality. This means that, for systems with post-editing, a mechanical print-reader will have to be used and most probably also a special purpose machine.

For the preparation of practical MT programs, great linguistic sophistication seems to be neither requisite nor even especially helpful at the present state of the art. Basic linguistic research is of great importance as such, and its support should preferably not be based on the pretense that it will lead to an improvement of MT techniques as is often done in the United States as well as in Russia. It is likely that far-reaching illumination of the human factor in translation will not be achieved without an enormous amount of such basic research, but this is a very long-range affair that should preferably be kept separate from immediate goals.

There has been a great amount of overlap in research among the various MT groups, not only among those in different countries but even within the same country. A certain amount of overlap is inevitable and even definitely helpful. It is my strong feeling, however, that the existing overlap is unnecessarily high and has led in many cases to costly repetitions of achievements as well as, and even more often so, of failures. This is not the place to offer recommendations for the improvement of this state of affairs.

Machine translation has not developed quite as speedily as its pioneers were hoping for seven or eight years ago, encouraged by spectacular initial successes. This is partly because the development of large-capacity and low-access-time memory devices has perhaps not quite fulfilled the high expectations of that time, and partly simply because of the vastness of the task which was not seen clearly enough at that time, so that by sticking too long to the goal of FAHQT much effort was wasted, at least insofar as immediate results are concerned. Nevertheless, commercial utilization of electronic machinery as aids in translation is now a practical prospect which will materialize after a series of additional improvements in linguistic and computer techniques along lines well understood at present. Speculations as to a breakthrough which will be made by the advent of learning machines, exciting as they are in themselves, have been left aside in this survey.

4 Remark on Bibliography

No attempt was made to provide here a complete bibliography. The journal *Mechanical Translation*, edited by W. N. Locke and V. H. Yngve of MIT, contains in addition to articles and news items an annotated bibliography. The last item in the bibliography of the last issue in my possession, Vol. 5, No. 1, dated July 1958 (published in December 1958), has the ordinal number 152.

Other bibliographies are given on pp. 227–236 of reference [17] (46 annotated items), pp. 82–95 of reference [7] (82 annotated items), pp. 22–51 of reference [2], and pp. 51–65 of reference [3] (contains some 170 items, including internal reports, work papers, etc.); Appendix 7 of reference [13] contains a complete bibliography of work performed up to the end of December 1958 at the Harvard Computation Laboratory on automatic translation and mathematical linguistics (58 items, mostly unpublished seminar papers). Useful current references are given *passim* in reference [4], and further references are undoubtedly contained in issue No. 4, of that survey which was scheduled to appear in April 1959.

Notes

This article was prepared with the sponsorship of the Information Systems Branch, Office of Naval Research, under Contract NR 049130. Reproduction as a whole or in part for the purposes of the U.S. Government is permitted.

1[6]. Among the exceptions should be mentioned the characterization (on p. 14) of the Polish logician Ajdukiewicz as a linguist and the similar mistake with regard to the Polish school of logicians. I myself am characterized on this occasion as the exponent of the Polish school in the United States, which is misleading in various ways. (It is true, however, that I acknowledged in a paper cited in reference [27] the impact of a certain article of Ajdukiewicz's which does not seem to have been read by the RAND group, though it appears in their bibliography.)

References

1. Bar-Hillel, Y., The present state of research on mechanical translation, *Am. Document.* 2, 229–237 (1951, appeared in 1953).

2. Edmundson, H. P., K. E. Harper, and D. G. Hays, Studies in machine translation—1: Survey and critique, Project RAND Research Memorandum RM-2063, 1958.

3. Reitwiesner, G. W., and M. H. Weik, Survey of the field of mechanical translation of languages, Ballistic Research Laboratories Memorandum Rept. No. 1147, 1958.

4. Science Information Service, Nat. Sci. Found., Current research and development in scientific documentation, No. 3, NSF-58-33, 1958.

5. Dostert, L., ed., *Research in Machine Translation*, Monograph Series on Languages and Linguistics No. 10, Georgetown Univ. Press, Washington, D.C., 1957.

6. Booth, A. D., L. Brandwood, and J. P. Cleave, *Mechanical Resolution of Linguistic Problems*. Academic Press, New York, 1958.

7. Appendix A, Machine Translation of Languages, Ramo-Wooldridge Corp. Design study for integrated USAF intelligence data handling system, p. 58, 1957.

8. See reference 2, p. 12.

9. See reference 7, p. 57.

10. Ramo-Wooldridge Project Prog. Rept. M20-8U13, Experimental machine translation of Russian to English, p. 5, 1958.

11. UNESCO Scientific and technical translating and other aspects of the language problem, documentation and terminology of science, pp. 103–110. Paris, 1957.

12. Oettinger, A. G., W. Foust, V. Giuliano, K. Magassy, and L. Matejka, Linguistic and machine methods for compiling and updating the Harvard Automatic Dictionary, to appear in the *Proc. Intern. Conf. Sci. Inform., Washington, D.C.*, pp. 137–l59 of Area 5 of preprinted vol. (1958).

13. Giuliano, V. E., *An Experimental Study of Automatic Language Translation*, Ph.D. Thesis, Harvard Univ., Cambridge, Mass., l959; Rept. No. NSF-l on mathematical linguistics and automatic translation, Harvard Univ. Computation Lab.

14. See reference 13, pp. 2–4.

15. See reference 5, p. 172.

16. Edmundson, H. P., and D. G. Hays, Studies in machine translation—2: Research methodology, Project RAND Research Memorandum RM-2060, 1957.

17. Weaver, W., Translation, in *Machine Translation of Languages* (W. N. Locke and A. D. Booth, eds.), pp. 15–23. Technology Press and Wiley, New York, 1955.

18. Reifler, E., The machine translation project at the University of Washington, Seattle, Washington, *Proc. 8th Intern. Congr. Linguists, Oslo*, pp. 514–518 (1958). This report forms part of a section meeting on machine translation, of which the proceedings are published on pp. 502–539 of this reference.

19. See reference 18, p. 514.

20. Shiner, G., The USAF Automatic Language Translator Mark I, 1958. *IRE Natl. Convention Rec.*, Part 4, pp. 296–301.

21. Reifler, E., Mechanical determination of the constituents of German substantive compounds, *Mech. Transl.* 2 (1), 3–14 (1955).

22. See reference 18, p. 517.

23. See reference 18, p. 517.

24. See reference 18, p. 518.

25. See reference 18, p. 516.

26. Yngve, V. H., The feasibility of machine searching of English texts, to appear in the *Proc. Intern. Conf. Sci. Inform., Washington, D.C.*, p. 167 of Area 5 of preprinted vol. (1958).

27. Chomsky, N., *Syntactic Structures*, Mouton, 's-Gravenhage, Holland, 1957 (with further references).

28. Bar-Hillel, Y., Decision procedures for structure in natural languages, *Logique et Analyse* [N. S.] 2 (5), 19–29 (1959).

29. Bar-Hillel, Y., A quasi-arithmetical notation for syntactic description, *Language* 29, 47–58 (1953).

30. Lambek, J., The mathematics of language structure, *Am. Math. Monthly* 65, 154–170 (1958).

31. Yngve, V. H., A programming language for mechanical translation, *Mech. Transl.* 5 (1), 25–41 (1958).

32. Garvin, P., Linguistic analysis and translation analysis, in *Research in Machine Translation* (L. E. Dostert, ed.), pp. 19–38. Georgetown Univ. Press, Washington, D.C., 1957.

33. Garvin, P., Syntactic units and operations, *Proc. 8th Intern. Congr. Linguists, Oslo*, pp. 626–632 (1958).

34. Lukjanow, A. W., Statement of proposed method for mechanical translation, Seminar Work Paper MT-35 of the Machine Translation Project of Georgetown Univ., 1957.

35. Zarechnak, M., Three levels of linguistic analysis in machine translation, *J. Assoc. Comp. Mach.* 6, 24–32 (1959).

36. Kaplan, A., An experimental study of ambiguity and context, The RAND Corporation, p. 187, 1949; published in *Mech. Transl.* 2 (2), 39–46 (1955).

37. See reference 7, p. 39.

38. Edmundson, H. P., and D. G. Hays, Research methodology for machine translation, *Mech. Transl.* 5 (1), 8–15 (1958) (revised version of reference 16).

39. Edmundson, H. P., K. E. Harper, D. G. Hays, and A. M. Koutsoudas, Studies in machine translation—9: Bibliography of Russian scientific articles, Project RAND Research Memorandum RM-2069, 1958.

40. Oettinger, A. G., *A Study for the Design of an Automatic Dictionary*, Ph.D. Thesis, Harvard Univ., Cambridge, Mass., 1954.

41. See reference 13.

42. Giuliano, V. E., The trial translator, an automatic programming system for the experimental machine translation of Russian to English, to appear in the *Proc. 1958 Eastern Joint Computer Conf.*, IRE and ACM.

43. Koutsoudas, A., Research in machine translation: I. General program; to be published.

44. Harris, Z. S., Linguistic transformations for information retrieval, to appear in the *Proc. Intern. Conf. Sci. Inform., Washington, D.C.*, pp. 124–136 of Area 5 of preprinted vol. (1958).

45. See reference 44, p. 127.

46. See reference 44, p. 131.

47. See reference 27, pp. 20 ff.

48. Lehmann, W. P., Structure of noun phrases in German. See reference 5, pp. 125–133.

49. See reference 4.

50. Oswald, V. A., Jr., The rationale of the idioglossary technique. See reference 5, pp. 63–69.

51. Richens, R. H., and A. D. Booth, Some methods of mechanized translation, in *Machine Translation of Languages* (W. N. Locke and A. D. Booth, eds.), Technology Press and Wiley, New York, 1955.

52. See reference 6.

53. This is the book cited in reference 51.

54. Booth, K. H. V., *Programming for an Automatic Digital Calculator*, Academic Press, New York, 1958.

55. See reference 6, p. 293.

56. Curry, H. B., and R. Feys, *Combinatory Logic*, North-Holland Publishing Co., Amsterdam, Holland, 1958.

57. See reference 29, p. 55.

58. Masterman, M., New techniques for analyzing sentence patterns (Abstr.), *Mech. Transl.* 3 (1), 4–5 (1956).

59. Masterman, M., R. M. Needham, and K. Sparck Jones, The analogy between mechanical translation and library retrieval, to appear in the *Proc. Intern. Conf. Sci. Inform., Washington, D.C.*, pp. 103–121 of Area 5 of preprinted vol. (1958).

60. Halliday, M. A. K., The linguistics of mechanical translation, *Proc. 8th Intern. Congr. Linguists, Oslo*, pp. 527–533 (1957).

61. Masterman, M., The thesaurus in syntax and semantics, *Mech. Transl.* 4 (1–2), 35–44 (1957).

62. King, G. W., A thesaurus lattice approach to the structure of language and communication with words, Report for the Natl. Sci. Found., 1958.

62a. See reference 6, p. 293.

63. Panov, D. Y., *Automatic Translation* (Russian), Popular Science Series, Academy of Sciences of the USSR, Moscow, 1958; tr. by U.S. *Joint Publications Research Service, JPRS 487-D*, pp. 1–20, see p. 2.

64. Belskaya, I. K., Machine translation of languages, *Research* (*London*) 10, 383–389 (1957).

65. Korolev, L., G. Rasoumovski, and G. Zelenkevitch, *Les Expériments de la Traduction Automatique de l'Anglais en Russe à l'Aide de la Calculatrice BESM*, Académie des Sciences de l'URSS, Moscou, 1956.

66. Mukhin, I. S., *An Experiment of the Machine Translation of Languages Carried Out on the BESM*, Academy of Sciences of the USSR, Moscow, 1956; also in *Proc. Inst. Elec. Engrs.* (*London*) 103, Part B, Suppl. 1–3, pp. 463–472 (1956), with minor editorial modifications.

67. Panov, D. Y., *Concerning the Problem of Machine Translation of Languages*, Academy of Sciences of the USSR, Moscow, 1956.

68. Rozentsveyg, V. Y., The work in the Soviet Union on machine translation from foreign languages into Russian and from Russian into foreign languages (Russian), *Repts. Intern. Congr. Slavicists, 4th Congr. Moscow*, 1958; tr. by L. R. Micklesen, Dept. of Far Eastern and Slavic Languages and Literatures, Univ. of Washington, Seattle, 1958.

69. See reference 63.

70. Nikolayeva, T. M., Russian sentence analysis (Russian), Rept. of ITMVT, Moscow, 1958; tr. by U.S. Joint Publications Research Service, JPRS/DC-387, pp. 6f.

71. See reference 70, p. 14.

72. Panov, D. Y., A. A. Lyapunov, and I. S. Mukhin, *Automatization of Translation from One Language to Another* (Russian), Academy of Sciences of the USSR, Moscow, 1956; tr. by U.S. Joint Publications Research Service, JPRS/DC 379, p. 25.

73. Fries, C. C., *The Structure of English*, Harcourt, Brace, New York, 1952.

74. Moloshnaya, T. N., Certain questions of syntax in connection with machine translation from English to Russian (Russian), *Vosprosy Yazvkoznaniya*, VI (4), 92–97 (1957); tr. by U.S. Joint Publications Research Service, JPRS/DC-68, pp. 48–57.

75. Kulagina, O. S., On a method for defining linguistic concepts (Russian), *Bull. Seminar on Problems of Machine Transl.*, No. 3, pp. 1–18 (1957).

76. Barkhudarov, L. S., and G. V. Kolshansky, The possibilities of machine translation (Russian), *Voprosy Yazykoznaniya*, VII (1), 129–133 (1958); tr. by U.S. Joint Publications Research Service, JPRS/DC-3I9, pp. 1–9.

77. Zhirkov, L. I., Limits of applicability of machine translation (Russian), *Voprosy Yazykoznaniya*, V (5), 121–124 (1956); tr. by U.S. Joint Publications Research Service, JPRS/DC-68, pp. 30–35.

78. Andreyev, N. D., Machine translation and the problem of an intermediate language (Russian), *Voprosy Yazykoznaniya*, VI (5), 117–121 (1957); tr. by U.S. Joint Publications Research Service, JPRS/DC-68, pp. 58–67.

79. Andreyev, N. D., ed., *Materials on Machine Translation* (Russian), Vol. I, Univ. of Leningrad Press, Leningrad, Russia, 1958.

80. Ivanov, V. V., Linguistic problems of constructing a machine language for information machines (Russian). See reference 79, pp. 10–39.

81. See reference 73.

82. Ceccato, S., La grammatica insegnata alle machine, *Civiltà delle Machine*, Nos. 1 and 2 (1956).

83. Bar-Hillel, Y., Report on the state of machine translation in the United States and Great Britain, Tech. Rept. No. 1, Hebrew Univ., Jerusalem, Israel, 1959.

[. . .]

Appendix I: MT Statistics as of April 1, 1959

(No responsibility as to the accuracy of the figures is undertaken. They were obtained by personal communication, the author's impressions or *bona fide* guesses. In cases of pure guesses, a question-mark is appended.)

Institution	Year of start of research	Number of workers	Full-time equivalents	Current yearly budget ($)	Project leader(s)
University of Washington Department of Far Eastern and Slavic Languages and Literature Seattle, Washington	1949	10?	6?	?	Erwin Reifler
Massachusetts Institute of Technology Research Laboratory of Electronics and Department of Modern Languages Cambridge 39, Massachusetts	1951	10?	6?	?	Victor H. Yngve
Georgetown University The Institute of Languages and Linguistics Machine Translation Project 1715 Massachusetts Avenue Washington, D.C.	1952	30?	15?	?	Leon E. Dostert Paul L. Garvin Ariadne W. Lukjanow Michael Zarechnak A. F. R. Brown
The RAND Corporation 1700 Main Street Santa Monica, California	(1950) 1957	15	9	?	David G. Hays Kenneth E. Harper
Harvard University The Computation Laboratory Machine Translation Project Cambridge 38, Massachusetts	1953	11	7?	?	Anthony G. Oettinger
University of Michigan Willow Run Laboratories Ann Arbor, Michigan	1955	11	7	?	Andreas Koutsoudas
University of Pennsylvania Department of Linguistics Philadelphia, Pennsylvania	1956?	10?	3?	?	Zellig S. Harris
National Bureau of Standards Washington, D.C.	1958	3	2	25,000	Ida Rhodes
Wayne State University Department of Slavic Languages and Computation Laboratory Detroit, Michigan	1958	10	6	40,000	Harry H. Josselson Arvid W. Jacobson
University of California Computer Center Berkeley, California	1958	8	5	40,500	Louis G. Henyey Sydney M. Lamb
University of Texas Department of Germanic Languages Austin 12, Texas	1958	?	?	?	Winfred P. Lehmann
Other American groups and individuals		50?	10?	?	
Total, USA		**150?**	**80?**	**1,500,000?**	
Birkbeck College Department of Numerical Automation London, England	(1947) 1955	6?	3?	?	Andrew D. Booth
Cambridge Language Research Unit 20 Millington Road Cambridge, England	1955?	20?	5?	?	Margaret Masterman
Institute of Precision Mechanics and Computer Engineering Academy of Sciences of the USSR Moscow	1955	?	?	?	I. S. Mukhin

Institution	Year of start of research	Number of workers	Full-time equivalents	Current yearly budget ($)	Project leader(s)
Steklov Mathematical Institute Academy of Sciences of the USSR 1 Akademichesky Proezd, Dom No. 28 Moscow V-134	1955	?	?	?	A. A. Lyapunov
Division of Applied Linguistics *Institute of Linguistics* Academy of Sciences of the USSR Moscow	1956?	?	?	?	A. A. Reformatsky I. A. Melchuk
Laboratory of Electrical Modelling All-Union Institute of Scientific and Technical Information Moscow	1956?	?	?	?	V. A. Uspensky
Experimental Laboratory of Machine Translation Leningrad State University Leningrad	1958	?	?	?	N. D. Andreyev
Association of Machine Translation Pedagogical Institute of Foreign Languages Moscow State University Moscow	1957?	?	?	?	I. I. Revzin V. Yu. Rozentsveyg V. V. Ivanov
Seminar on Mathematical Linguistics Department of Philology Moscow State University Moscow	1956	?	?	?	P. S. Kuznetsov
Other Russian groups and individuals		?	?	?	
Total, Russia		**300?**	**120?**	**1,500,000?**	
University of Milan Milan, Italy	1958	?	?	?	Silvio Ceccato
Hebrew University Jerusalem, Israel	1958	4	1	4,000	Yehoshua Bar-Hillel
Other groups and individuals		?	?	?	
Total		**500?**	**220?**	**3,000,000?**	

Appendix II: Some Linguistic Obstacles to Machine Translation*

For certain pairs of languages it has been shown experimentally that word-by-word machine translation leads to an output which can often be transformed by an expert post-editor into a passable translation of the source text. However, if one is interested in reducing the burden of the post-editor, as one apparently has to be in order to make the use of machine aids in translation commercially profitable, or if one has to do with pairs of languages for which word-by-word translation is not by itself a satisfactory basis for post-editing, it is natural to think of mechanizing the determination of the syntactic structure of the source sentences. It is theoretically clear, and has again been experimentally verified, that knowledge of the syntactic structure of the sentences to be translated does considerably simplify the task of the post-editor. It is obvious, for instance, that this knowledge tends to reduce, and in the limit to eliminate, those syntactical ambiguities which are created by the word-by-word translation and which are nonexistent for the human translator who treats the sentences as wholes. The task of the post-editor would then consist solely in eliminating the semantical ambiguities and in polishing up the style of the machine output. Whether these steps, too, can be completely taken over by machines of today or of the foreseeable future is still controversial. I myself have strong reasons for regarding it

* This appendix comprises greatly revised versions of parts of two talks, one given before the Colloque de Logique, Louvain, September 1958 and published in *Logique et Analyse* 2, 19–29 (1959), the other given before the Second International Congress of Cybernetics, Namur, September 1958, to be published in the Proceedings of this congress, 1960.

as hopeless, in general, but this is not the point I would like to discuss here; it is the subject of Appendix III.

A few years ago, I proposed what I called a *quasi-arithmetical notation for syntactic description* [1] whose employment should allow, after some refinements, for a mechanical determination of the constituent structure of any given sentence. At that time, I actually demonstrated the effectiveness of the method for relatively simple sentences only, but cherished the hope that it might also work for more complex sentences, perhaps for all kinds of sentences. I am now quite convinced that this hope will not come true. As a consequence, the road to machine translation can be shown to contain more obstacles than was realized a few years ago. I think that this should be of sufficient interest to warrant some more detailed exhibition, especially since this insight is due to an important new, not to say revolutionary, view of the structure of language, recently outlined by the American linguist and logician Noam Chomsky [2], and could perhaps, in its turn and in due time, be turned into a new method of machine translation, which would be more complex than the known ones but also more effective.

Since I can not assume acquaintance with the paper in which I introduced the quasi-arithmetical syntactical notation mentioned above, let me present its main point here again very briefly, with some slight modifications in terminology and notation, partly under the impact of a recent article of Lambek [3]; for a full presentation, the paper should be consulted.

The basic idea, adopted from a paper of the Polish logician Ajdukiewicz [4], is to regard every sentence (of more than one word) as the result of the operation of one continuous part of it upon the remainder, these two parts being the *immediate constituents* of the sentence, such that these constituent parts which in general are not sentences themselves, but rather phrases, are again the product of the operation of some continuous part upon the remainder, etc., until one arrives at the final constituents, say words or morphemes. In accordance with this variant of the *immediate constituent model*, which is the standard model with which many modern linguists are working [5], all words of a given language are assigned to one or more, but always finitely many, *syntactic categories*.

For the purpose of illustration we shall try to get along, for English, with two fundamental categories, those of *nominals* and (declarative) *sentences*, to be denoted by n and s, respectively. The operator category of *intransitive verbals*, i.e., the category of those words that out of a nominal to their left form a sentence, will be denoted by $n\backslash s$ (read: n sub s), the category of *adjectivals*, i.e., of words that out of nominals to their right form nominals, will be denoted by n/n (read: n super n), the category of *intransitive verbal adverbals*, i.e., of words that out of intransitive verbals (to their left) form intransitive verbals, by $(n\backslash s)\backslash(n\backslash s)$—for which we shall, by means of a self-explanatory convention, usually write $n\backslash s\backslash\backslash n\backslash s$—etc. (Nominals, verbals, adjectivals, etc., in my present usage, are *syntactical categories*. They should not be confused with nouns, verbs, adjectives, etc., which are *morphological* (*paradigmatic*) *categories*, in my usage. The connection between these two classifications, as the choice of terms is intended to indicate, is that nouns usually, though by no means always, belong to the syntactical category of nominals, etc., and that most expressions belonging to the syntactical category of nominals, of course only if they are single words, are nouns.) In *Little John slept soundly*, for instance, we would regard *Little* to be an n/n, *John* an n, *slept* an $n\backslash s$, and *soundly* an $n\backslash s\backslash\backslash n\backslash s$.

Assuming then, that a category "dictionary" listing for each English word all its categories stands at our disposal, the task of finding out whether a given word sequence is a sentence or, more generally, a *well-formed* (or *connex*) expression and, if so, what its *constituent structure* is, could now be solved according to the following utterly mechanical procedure: We would write under each word of the given word sequence the symbols for all the categories to which it belongs, separated by commas, and then start *cancelling* in all possible ways, according to either of the two following rules:

$$\alpha, \alpha\backslash\beta \rightarrow \beta \quad \text{and} \quad \alpha/\beta, \beta \rightarrow \alpha.$$

(The reading of these rules should be self-explanatory. The first, for instance, reads: Replace the sequence of two category symbols, the first of which is any category symbol whatsover and the second of which consists of the first symbol followed by a left diagonal stroke followed by any category symbol whatsoever, by this last category symbol.) A series of such symbol sequences where each sequence results from its predecessor by one application of a cancellation rule is called a *derivation*. The last line of a derivation is its *exponent*. If the exponent consists of a single symbol, *simple* when it consists of a single letter, *complex* when it contains at least one stroke, the word sequence with this exponent, and with the constituent structure given by the derivation, is well-formed; if the exponent of a certain derivation

is, more specifically, s, the sequence is a sentence, relative to this derivation.

To illustrate, let us start with the last analyzed expression:

Little John slept soundly.

Let us assume (contrary to fact) that as a result of consulting the category dictionary we would have arrived at just the following category symbol sequence:

(1) $n/n, \quad n, \quad n\backslash s, \quad n\backslash s\backslash\backslash n\backslash s.$

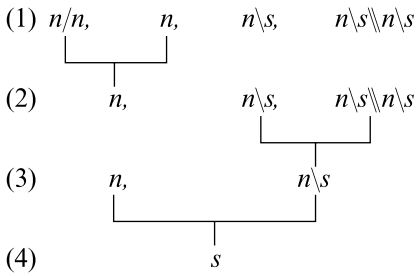
It is easy to see that there are exactly three different ways of performing the first cancellation, starting off three different derivations, viz.:

(2) $n, \quad n\backslash s, \quad n\backslash s\backslash\backslash n\backslash s.$

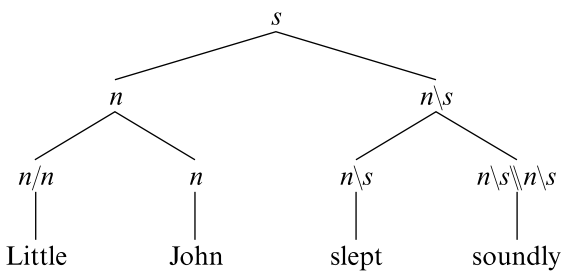
(2′) $n/n, \quad s, \quad n\backslash s\backslash\backslash n\backslash s.$

(2″) $n/n, \quad n, \quad n\backslash s.$

(2′) leads into a blind alley. The other two lines, (2) and (2″), each allow for two continuations, of which one again leads into a blind alley, whereas the other allows for just one more derivation, with both exponents being s. Let me write down one of these derivations:



The other derivation differs from the one just presented only in that the two cancellation steps in (2) and (3) occur in the opposite order. These two derivations are therefore *equivalent* in an important sense; in fact, they correspond both to the same *tree expansion:*



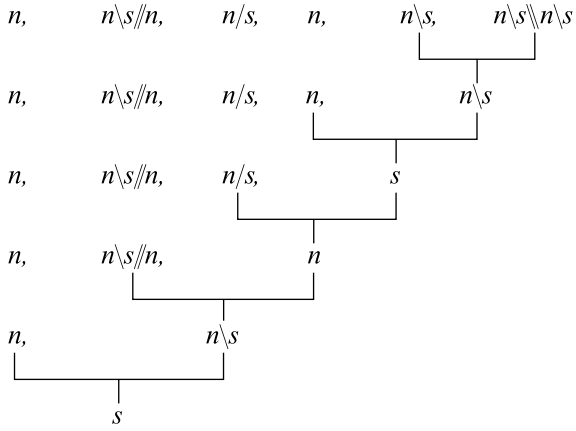
Our second and final example will be:

Paul thought that John slept soundly.

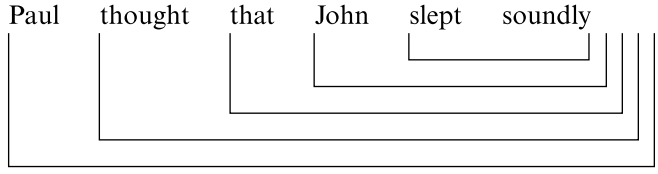
(I hope that the somewhat shaky English of this example will be forgiven; it simplifies making the point without falsifying it.) Copying only the first entry under each word in our fictitious category dictionary, we arrive at

Paul	thought	that	John	slept	soundly
$n,$	$n\backslash s//n,$	$n/s,$	$n,$	$n\backslash s,$	$n\backslash s\backslash\backslash n\backslash s$

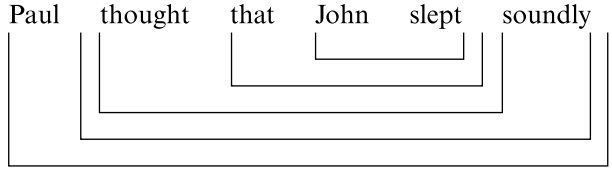
There are two nonequivalent derivations with a single exponent. I shall again write down only one of these derivations:



The constituent structure corresponding to this derivation can be pictured in the following parsing diagram:



The reader is invited to check that the parsing diagram corresponding to the other derivation is:

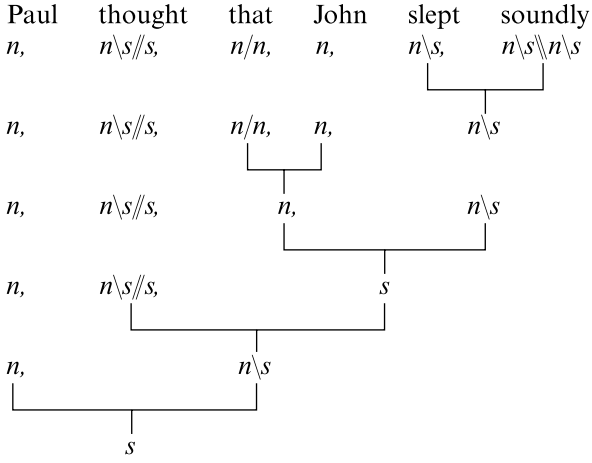


If this structure is regarded as unacceptable, this would prove that either the categorization assumed for this illustration is ill-chosen, or else that the whole model is inadequate for English. I shall not pursue this issue further here, since I shall later on proffer stronger reasons for questioning the adequacy of the model.

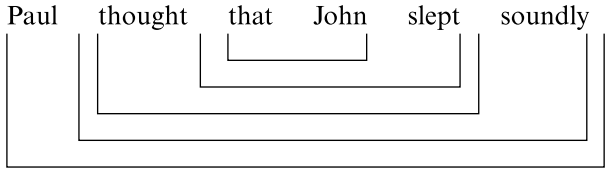
As said before, the situation actually is more complicated. An adequate category dictionary would contain in general more than one entry per word. *That*, e.g., is often a nominal, n, and even more often an adjectival, n/n; *soundly* could as well be an $n\backslash s /\!/ n\backslash s$ or an $(n\backslash s)/n /\!/ (n\backslash s)/n$, and *thought*, finally, belongs also to categories n, $n\backslash s$, $n\backslash s /\!/ s$ (*Paul thought John was asleep*) and, as a participle, to still others. It can nevertheless readily be seen that our method is capable, at least in certain cases, to determine by purely mechanical operations the specific category to which a given word belongs in its given linguistic context. In our example, e.g., listing all the mentioned categories in column form yields the following scheme:

Paul	thought	that	John	slept	soundly
n	$n\backslash s /\!/ n$	n/s	n	$n\backslash s$	$n\backslash s \backslash\!\backslash n\backslash s$
	n	n			$n\backslash s /\!/ n\backslash s$
	n/s	n/n			$(n\backslash s)/n /\!/ (n\backslash s)/n$
	$n\backslash s /\!/ s$				

It would be a tedious but wholly routine exercise to determine that out of the very many derivations corresponding to this word sequence—notice that there are 36 initial lines alone!—there exist only four essentially different ones with a single exponent, namely, in addition to the two above-mentioned derivations, just

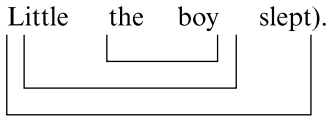


and the one whose parsing diagram would be



I still remember my surprise a few years ago when I discovered that the third constituent structure is doubtless grammatical, however wildly implausible the conditions under which it would be uttered. I was still less pleasantly surprised when I discovered recently that the second and fourth derivations yield constituent structures whose grammaticalness is highly doubtful. The sentence under discussion is *syntactically ambiguous* (or *constructionally homonymous*).

So far, so good then. But, unfortunately, the actual situation is still much more complicated. For categorizing English words, it would be necessary to increase the number of fundamental categories and distinguish various kinds of nominals, for instance, *singular* and *plural*, *animate* and *inanimate*. One would also have to distinguish *declarative sentences*, *question sentences*, etc., these being irreducible to each other under the present model. Some additional notational means would have to be found from which it will follow that *John slept*, *The boy slept*, *Boys slept*, *The boys slept* are well-formed but that *Boy slept*, *The John slept* are not, that *The little boy slept* is connex but not *Little the boy slept* (at least not with the parsing diagram



These and thousands of other additional refinements could perhaps still be introduced without blowing up the whole model. But there are many features which make it highly doubtful whether English grammar—or that of any other natural language, for that matter—can at all be forced into the straitjacket of the immediate constituent model and remain workable and revealing. Since the arguments against such a possibility have already been presented by Chomsky [6], I shall not repeat them here in all their generality but restrict myself to the point of view of machine translation.

It takes but little to realize that the four categories mentioned above for *thought* are far from being exhaustive. In addition to its being a participle, which has already been mentioned, these are such phrases as *thought processes*, *thought thirsty* (not common but definitely grammatical), *thought provoking*, etc. In order to take care of the first two contexts, for example, we would have to assign *thought* also to the categories n/n and $n/n /\!/ n/n$. (“In these contexts, *thought* occurs in the function of an adjective or an

adverb, respectively'' would have been one traditional way of putting the issue.) The third context would have raised the notoriously difficult problem of the status of the participle present, in addition. The task of preparing a category list that would work for all these and many other contexts is certainly much harder than the first successful analyses caused us to believe. Would not the required list become so long that the mechanical determination of the constituent structure of, say, a 30-word sentence might well require trillions of machine operations, hence be totally impractical for machines of today as well as of tomorrow?

It is likely, for instance, that every assignment to a category of the form $\alpha\backslash\beta/\!/\gamma$ (such as the assignment of *thought* to $n\backslash s/\!/s$) will have to be accompanied by an assignment to the category $\alpha\backslash\!\backslash\beta/\gamma$; an attempt to avoid this through a unique assignment to $\alpha\backslash\beta/\gamma$ would amount to a change in the notational framework and would require considerable changes in the cancellation rules.

Still worse, it is not clear whether assigning each word to a finite number of categories only would do at all. Should it be requisite to assign some words to an infinity of categories, then the simple mechanical procedure described above of determining the syntactic structure of a given word sequence breaks down, since there is no longer any assurance that the number of derivations is finite at all.

And what about a sentence such as *Playing cards is fun?* On first sight, it seems that one has to arrive at the category n for the phrase *playing cards*. However, it is intuitively clear that this should not be derived from *cards* being an n and playing being an n/n (and not only intuitively so: notice that the next word is *is* and not *are*; *playing cards* in our context is a singular nominal). There are, of course, many other ways of enforcing an assignment of n to *playing cards*, but none of these, to my knowledge, is such that it would not introduce unwarranted and counter-intuitive syntactical resolutions of other sentences. ''Hocus-pocus'' linguistics—as certain linguistic methods were called whose only purpose was to save certain phenomena, without regard to any intuitive (or psychological) realities—would in our case definitely refute itself by saving also phenomena that are nonexisting.

And what about a sentence like *He looked it up*? We all feel that *looked* and *up* belong together and that in the context *He looked up the table*, at any rate, *up* is an operator that out of an intransitive verbal to its left forms a transitive verbal, hence belongs to the category $n\backslash s\backslash\!\backslash(n\backslash s)/n$. This assignment indeed works well for *He looked up the table*, but would obviously not do for *He looked it up*, since there exists no derivation from

$$n, \quad n\backslash s, \quad n, \quad n\backslash s/\!/(n\backslash s)/n$$

with an exponent s.

Finally, what about a sentence like *John, unfortunately, was asleep*? *Unfortunately*, in the context *Unfortunately, John was asleep*, is clearly an s/s but with this category assignment, *John, unfortunately, was asleep* would turn out not to be connex.

If now the present variant of an immediate constituent model is not good enough to serve as a general model for the whole syntax of a given language, the method of mechanical structure determination outlined above can no longer be assumed to be of general validity, either. As a matter of fact, I had already noticed six years ago that the model did not work too well for complex sentences, but had rather hoped that this was due only to lack of refinement that could be partly remedied by increasing the number of fundamental categories, partly by using additional rules. I have now come to realize that its failure in the more complex cases has a much deeper cause: the linguistic model on which this method was based is just not good enough.

The situation is apparently not changed very much by using a more complex model which has recently been proposed by Lambek [7]. Though he uses in addition to the cancellation rules other rules of a different character which may perhaps allow for a reduction in the number of the machine operations required for a test of sentencehood, it is not clear whether Lambek's model is really more powerful than the one outlined above.

Another model, or rather a whole set of models, for linguistic structure has recently been developed, in outline, by Chomsky [8, 9, 10]. (A similar conception has been developed also by Harris [11], but since Chomsky's formulations seem to me much clearer, I prefer to refer to his work in the sequel.) They are incomparably more powerful than the phrase structure models, in all their variants. These so-called *transformational models* do not discard the immediate constituent model but rather supplement it. The former model remains intact for a certain kind of simple sentences, the so-called *kernel sentences* (or rather for their underlying *terminal strings*)—and our method of mechanical structure determination remains therefore valid for these sentences—but has to be supplemented

by additional procedures, the so-called *transformations*, in order to account for the synthesis of *all* sentences.

Each sentence, according to the transformational models, is the result of a series of one or more transformations performed one after the other on one or more terminal strings—unless, of course, it is a terminal string itself. A complete analysis, mechanical or otherwise, of a given sentence has to tell us what its basic terminal strings are, together with their constituent structure, and what transformations, and in what order, were performed upon them. Assuming that a complete transformational grammar for some given language has been prepared, the preparation of a corresponding analytical (or operational) grammar is a formidable, though perhaps not necessarily impossible task. There exist here a large number of unsolved problems, partly due to the fact that the nature of the transformations involved have so far been left rather vague, partly to the fact that we find ourselves here within the confines of new and extremely complicated disciplines like recursive function theory, Post canonical systems and the like, the exploration of which has only started. So far, of course, no transformational grammar exists for any language, to any serious degree of completeness.

The recognition that immediate constituent grammars have to be supplemented by transformational grammars makes the task of mechanizing translation look much harder, but the resulting picture is not at all uniformly black. On the contrary, there are reasons to suppose that the additional insight we get on the basis of this model will not only be of decisive importance for theoretical linguistics, but may well turn out to facilitate the mechanization of translation from new angles.

One gain of the transformational model is similar to, but still more effective and more intuitive than the one obtained by Lambek's model: a reduction in the number of categories to which the words will have to be assigned. No longer will *thought* have to be assigned to the categories n/n and $n/n/\!/n/n$ in order to take care of the connexity of *thought processes* and *thought thirsty*, because sentences containing these phrases are not terminal strings but result from transformations. In addition, the assignment of *thought* to $n\backslash s/\!/s$ is no longer required, since the sentencehood of *Paul thought John slept soundly* will now be taken care of by our regarding it as the result of a *that*-omitting transformation on *Paul thought that John slept soundly*, which itself is the result of a certain fusing transformation on the two terminal strings *Paul thought this*. *John slept soundly*. (This description is oversimplified and to that degree misleading. A better description is given in Chomsky's publications. A sufficiently sophisticated treatment would require too much space here.) As a result, the noun *thought* will (perhaps) always be assigned to the syntactical category of nominals, the finite verb *thought* to the syntactic category of transitive verbals, and the participle *thought* to an appropriate syntactical category (with which we shall not bother here), the multiplicity of category assignments to the word *thought* now being considered as exclusively the result of homonymity or homography, as the case may be.

One interesting result of all this will now be that the number of categories of many words will be reduced to—zero. This will happen if no sentence containing these words is regarded as a terminal string. To give an example: *sleeping* will not be assigned to any category, any sentence containing this word being considered as the result of a transformation. (*Interesting*, however, will be assigned to the category n/n, the difference being—to give only a hint—that *very interesting* is connex but not *very sleeping*.) That there might be words which do not belong to any syntactic category will strike many linguists as rather queer, but I am convinced that on second sight they will realize the enormous advantages of such an attitude; innumerable pseudo-problems have in the past been created by the search for the syntactic category (the traditional term is, of course, "part of speech") of certain words or phrases which—under the new model—just do not belong to any category. This is—if I may be allowed one generalization—just one more instance of the very common class of situations where the attempt of applying a model which is very useful within certain limits leads, when pushed beyond these limits, to pseudo-problems and their pseudo-solutions.

The second gain is somewhat more speculative: it seems likely, but has so far not been seriously tested, that languages will be much more similar with regard to their terminal string structure than with regard to the structure of the totality of their sentences. Word-by-word translation of terminal strings, with some occasional permuting, seems to yield satisfactory results for many pairs of languages, including those for which this kind of translation does not work at all with regard to more complex sentences.

The most remarkable gain, however, would be achieved when it turned out that between the sets of transformations of two languages there existed a close semantic relationship. Should it happen that for

certain two languages, L_1 and L_2, there exist two transformations, say t_1 and t_2, such that for any semantically equivalent terminal strings of these languages, k_1 and k_2, $t_1(k_1)$ is semantically equivalent to $t_2(k_2)$, this would allow for a relatively simple mechanization of the translation, provided, of course, that the syntactic analysis of L_1 has been mechanized, whereas a word-by-word translation of $t_1(k_1)$ into L_2 might be highly unsatisfactory.

Of course, there is but little hope that the sets of transformations of two languages which do not stand in any close genetical relationship will do us the favor of exhibiting isomorphism or near-isomorphism with regard to semantic equivalence. So far, there exists to my knowledge no *general* theory of machine translation which would ensure that, if only the precepts of this theory are followed, the target-language counterpart (or counterparts) of any sentence of a given source-language will be no more and no less syntactically ambiguous than the original sentence itself. Current statements to the contrary seem to me palpably false, and any hope for an imminent establishment of such a theory—unsubstantiated. Great progress has been made in this respect with regard to certain ordered pairs of languages, such as French–English, German–English, Russian–English, English–Russian, German–Russian, and French–Russian, partly prior to the appearance of the transformational model and without any conscious use of its methods, and more progress may be expected in the future through a conscious use of these methods. As one almost necessary condition for future success I regard the recognition on behalf of the workers on machine translation that the model with which they were working, consciously or unconsciously, during the first decade of their endeavors was too crude and has to be replaced by a much more complex but also much better fitting model of linguistic structure.

References for Appendix II

1. Bar-Hillel, Y., A quasi-arithmetical notation for syntactic description, *Language* 29, 47–58 (1953).

2. Chomsky, N., *Syntactic Structures*, Mouton, 's-Gravenhage, Holland, 1957.

3. Lambek, J., The mathematics of sentence structure, *Am. Math. Monthly* 65, 154–170 (1958).

4. Ajdukiewicz, K., Die syntaktische Konnexitaet, *Studia Philosophica* 1, 1–27 (1935).

5. Hockett, C. F., *A Course in Modern Linguistics*, Macmillan, New York, 1958, especially section 17.

6. See reference 2.

7. See reference 3.

8. See reference 2.

9. Chomsky, N., Three models for the description of language, *IRE Trans. on Inform. Theory*, IT-2, No. 3 (1956).

10. Chomsky, N., A transformational approach to syntax; forthcoming.

11. Harris, Z. S., Co-occurrence and transformation in linguistic structure, *Language* 33, 283–340 (1957).

12. See reference 2, p. 45.

Appendix III: A Demonstration of the Nonfeasibility of Fully Automatic High Quality Translation

One of the reasons why we do not as yet have any translation centers, not even in the planning stage, in which electronic computers, general or special purpose, are used to automate certain parts of the translation process, in spite of the fact that such centers would fulfill a vital function in saving a considerable amount of qualified human translator time per document translated, and thereby facilitate more, quicker and, after some time, cheaper translation, is the reluctance of many MT workers to recognize that the idea of inventing a method for fully automatic high quality translation (FAHQT) is just a dream which will not come true in the foreseeable future. By not realizing the practical futility of this aim, whatever its motivational importance for certain types of basic research, they have misled themselves and the agencies which sponsored their research into not being satisfied with a partly automated translation system whose principles are well understood today, and instead to wait for the real thing which was believed, and made to believe, to be just around the corner.

During the past year I have repeatedly tried to point out the illusory character of the FAHQT ideal even in respect to mechanical determination of the syntactical structure of a given source-language sentence. [...] Here I shall show that there exist extremely simple sentences in English—and the same holds, I am sure, for any other natural language—which, within certain linguistic contexts, would be uniquely (up to plain synonymy) and unambiguously translated into any other language by anyone with a sufficient knowledge of the two languages involved, though I know of no program that would enable a machine to come up with this unique rendering unless by a completely arbitrary and *ad hoc* procedure whose futility would show itself in the next example. A sentence of this kind is the following:

The box was in the pen.

The linguistic context from which this sentence taken is, say, the following:

Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.

Assume, for simplicity's sake, that *pen* in English has only the following two meanings: (1) a certain writing utensil, (2) an enclosure where small children can play. I now claim that no existing or imaginable program will enable an electronic computer to determine that the word *pen* in the given sentence within the given context has the second of the above meanings, whereas every reader with a sufficient knowledge of English will do this "automatically." Incidentally, we realize that the issue is not one that concerns translation proper, i.e., the transition from one language to another, but a preliminary stage of this process, or, the determination of the specific meaning in context of a word which, in isolation, is semantically ambiguous (relative to a given target-language, if one wants to guard oneself against the conceivable though extremely unlikely case that the target-language contains a word denoting both the same writing utensil and an enclosure where children can play).

It is an old prejudice, but nevertheless a prejudice, that taking into consideration a sufficiently large linguistic environment as such will suffice to reduce the semantical ambiguity of a given word. Let me quote from the memorandum which Warren Weaver sent on July 15, 1949 to some two hundred of his acquaintances and which became one of the prime movers of MT research in general and directly initiated the well known researches of Reifler and Kaplan [1]: "... if ... one can see not only the central word in question, but also say N words on either side, then, if N is large enough one can *unambiguously* [my italics] decide the meaning of the central word. The formal truth of this statement becomes clear when one mentions that the middle word of a whole article or a whole book is unambiguous if one has read the whole article or book, providing of course that the article or book is sufficiently well written to communicate at all." Weaver then goes on to pose the practical question: "What minimum value of N will, at least in a tolerable fraction of cases, lead to the correct choice of meaning for the central word," a question which was, we recall, so successfully answered by Kaplan. But Weaver's seemingly lucid argument is riddled with a fateful fallacy: the argument is doubtless valid (fortified, as it is, by the escape clause beginning with "providing") but only for *intelligent* readers, for whom the article or book was written to begin with. Weaver himself thought at that time that the argument is valid also for an electronic computer, though he did not say so explicitly in the quoted passage, and on the contrary, used the word "one"; that this is so will be clear to anyone who reads with care the whole section headed "Meaning and Context." In this fallacious transfer Weaver has been followed by almost every author on MT problems, including many Russian ones. Now, what exactly is going on here? Why is it that a machine with a memory capacity sufficient to deal with a whole paragraph at a time, and a syntactico-semantic program that goes, if necessary, beyond the boundaries of single sentences up to a whole paragraph (and, for the sake of the argument, up to a whole book)—something which has so far not gotten beyond the barest and vaguest outlines—still powerless to determine the meaning of *pen* in our sample sentence within the given paragraph? The explanation is extremely simple, and it is nothing short of amazing that, to my knowledge, this point has never been made before, in the context of MT, though it must surely have been made many times in other contexts. What makes an intelligent human reader grasp this meaning so unhesitatingly is, in addition to all the other features that have been discussed by MT workers (Dostert [2], e.g., lists no less than seven of what he calls areas of meaning determination, none of which, however, takes care of our simple example), his *knowledge* that the relative sizes of pens, in the sense of writing implements, toy boxes, and pens, in the sense of playpens, are such that when someone writes under ordinary circumstances and in something like the given context, "The box was in the pen," he almost certainly refers to a playpen and most certainly not to a writing pen. (The occurrence of this sentence in the mentioned paragraph tends to increase the confidence of the reader that the circumstances are ordinary, though the whole paragraph could, of course, still have formed part of a larger fairy tale, or of some dream story, etc.) This knowledge is not at the disposal of the electronic computer and none of the dictionaries or programs for the elimination of polysemy puts this knowledge at its disposal.

Whenever I offered this argument to one of my colleagues working on MT, their first reaction was: "But why not envisage a system which will put this knowledge at the disposal of the translation machine?" Understandable as this reaction is, it is very easy to show its futility. What such a suggestion amounts to, if taken seriously, is the requirement that

a translation machine should not only be supplied with a dictionary but also with a universal encyclopedia. This is surely utterly chimerical and hardly deserves any further discussion. Since, however, the idea of a machine with encyclopedic knowledge has popped up also on other occasions, let me add a few words on this topic. The number of facts we human beings know is, in a certain very pregnant sense, infinite. Knowing, for instance, that at a certain moment there are exactly eight chairs in a certain room, we also know that there are more than five chairs, less than 9, 10, 11, 12, and so on *ad infinitum*, chairs in that room. We know all these additional facts by inferences which we are able to perform, at least in this particular case, instantaneously, and it is clear that they are not, in any serious sense, stored in our memory. Though one could envisage that a machine would be capable of performing the same inferences, there exists so far no serious proposal for a scheme that would make a machine perform such inferences in the same or similar circumstances under which an intelligent human being would perform them. Though a lot of thought should surely be given to the problems which could only be touched slightly here, it would very definitely mean putting the horse before the cart if practical MT would have to wait for their solution. These problems are clearly many orders of magnitude more difficult than the problem of establishing practical machine aids to translation. I believe that it is of decisive importance to get a clear view of this whole issue and hope that my remarks will contribute to its clarification.

I have no idea how often sentences of the mentioned kind, whose ambiguity is resolvable only on the basis of extra-linguistic knowledge which cannot be presumed to be at the disposal of a computer, occur on the average in the various types of documents in whose translation one might be interested. I am quite ready to assume that they would occur rather infrequently in certain scientific texts. I am ready to admit that none might occur on a whole page or even in some whole article. But so long as they will occur *sometimes*, a translation outfit that will claim that its output is of a quality comparable to that of a qualified human translator will have to use a post-editor, and this not only for polishing up purposes, contrary to what even so acute and impartial an observer as Warren Weaver was still hoping for in 1955 [3]. As soon as this is granted, the greatest obstacle to practical MT has been overcome, and the way is free for an unprejudiced discussion of the best human use of the human partner in the translation outfit.

Having shown, I hope, that FAHQT is out of the question for the foreseeable future because of the existence of a large number of sentences the determination of whose meaning, unambiguous for a human reader, is beyond the reach of machines, let me now discuss this issue of reduction of semantical ambiguity a little further. There exist in the main two methods of reducing semantical ambiguity. One is the use of idioglossaries, the other is the already mentioned method of utilizing the immediate linguistic environment of the word which is ambiguous in isolation. Though some doubts have been raised on occasion as to the validity of the first of these methods, I do not know of any serious attempt to put its validity to test. At this point I would only like to stress the vital necessity of performing such tests before an MT method based upon the utilization of idioglossaries is claimed to yield high quality translations, even in collaboration with a post-editor. It is just the great effectiveness of the use of idioglossaries in general which is apt to yield disastrously wrong translations on occasion without giving the post-editor even a chance to correct these mistakes. It is just because a certain Russian word in a chemical paper will *almost always* have a certain specific English rendering that the danger is so great that in those exceptional cases where this word, for some reason or other, will have a different meaning, this exception will not be taken into account, yielding a meaningful but wrong translation.

In regard to the second method, the situation is even worse, and has lately become even more confused through the use of certain slogan terms like "thesaurus" in this connection. (Notice, e.g., that the very same—fictitious!—thesaurus approach for English-to-French translation that would correctly render *pen* by "plume" in the sentence *The pen was in the inkstand* would incorrectly render *pen* by "plume" in the sentence *The inkstand was in the pen*.) It is undoubtedly true that consideration of the immediate linguistic neighborhood of a given ambiguous word is a very powerful method, but it is again necessary to realize its limitations. I am referring no longer to those limitations which I pointed out through the use of my sample sentence, but rather to the fact that many MT workers seem to underestimate the importance of those cases of reduction of polysemy which cannot be obtained by looking at the immediate neighborhood, and even more so about the fact that partial successes in this direction have led many people to underestimate the depth of the remaining gap. Let me state rather dogmatically that there exists at this moment no method of reducing the polysemy of

the, say, twenty words of an average Russian sentence in a scientific article below a remainder of, I would estimate, at least five or six words with multiple English renderings, which would not seriously endanger the quality of the machine output. It is looking at the quantities involved which creates a distorted picture with many people. Many tend to believe that by reducing the number of initially possible renderings of a twenty-word Russian sentence from a few tens of thousands (which is the approximate number resulting from the assumption that each of the twenty Russian words has two renderings on the average, while seven or eight of them have only one rendering) to some eighty (which would be the number of renderings on the assumption that sixteen words are uniquely rendered and four have three renderings apiece, forgetting now about all the other aspects such as change of word order, etc.) the main bulk of this kind of work has been achieved, the remainder requiring only some slight additional effort. We have before us another case of what, in a superficially different but intrinsically very similar situation, has been called the "80% fallacy" [4]. The remaining 20% will require not one quarter of the effort spent for the first 80%, but many, many times this effort, with a few percent remaining beyond the reach of every conceivable effort.

References for Appendix III

1. This memorandum is reprinted as chapter 1 of *Machine Translation of Languages* (W. N. Locke and A. D. Booth, eds.), Wiley, New York, 1955. The quoted passage appears there on p. 21. For Reifler's and Kaplan's studies, see p. 227 of the same volume.

2. Dostert, L. E., The Georgetown–IBM experiment, in *Machine Translation of Languages* (W. N. Locke and A. D. Booth, eds.), chapter 8, especially pp. 129 ff.

3. In *Machine Translation of Languages* (W. N. Locke and A. D. Booth, eds.), p. VII.

4. Bull, W. E., C. Africa, and D. Teichroew, Some Problems of the "word," in *Machine Translation of Languages* (W. N. Locke and A. D. Booth, eds.), chapter 5, p. 98.

7
A New Approach to the Mechanical Syntactic Analysis of Russian

Ida Rhodes

This chapter categorically rejects the possibility of considering a word-to-word conversion as a translation. A true translation is unattainable, even by the human agent, let alone by mechanical means. However, a crude practical translation is probably achievable. The present chapter deals with a scheme for the syntactic integration of Russian sentences.

Introduction

From the moment that a writer conceives an idea which he desires to communicate to his fellow men, sizable stumbling blocks are strewn in the path of the future translator. For the ability to shape one's thought clearly, or even completely, is not granted to many; rarer still is the gift of expressing the thought—precisely, concisely, unambiguously—in the form of words. There is no guarantee, therefore, that the author's written text is a reliable image of his original idea.

Furnished with this more or less distorted record, the translator is expected to perform a number of amazing feats. In the first place, he has to discern—often through the dim mist of the source language—the writer's precise intention. This requires not only a perfect knowledge of both the source language and the subject matter treated in the text, but also the mental skills customarily exercised by the professional sleuth. In addition, these newly reconstructed ideas must be rendered into a target language which is so unequivocal—and so faithful to the source—as to convey, to every reader of the translator's product, the exact meaning of the original foreign text!

Small wonder, then, that a fabulous achievement like Fitzgerald's translation of the Rubaiyat is regarded in the nature of a miracle. For the general case, it would seem that characterizing a sample of the translator's art as a *good* translation is akin to characterizing a case of mayhem as a *good* crime: in both instances the adjective is incongruous.

If, as a crowning handicap, we are asked to replace the vast capacity of the human brain by the paltry contents of an electronic contraption, the absurdity of aiming at anything higher than a crude *practical* translation becomes eminently patent.

Perhaps we are belaboring this point; we do so to avoid later arguments about the "quality" of our work. If, for example, a translated article enables a scientist to reproduce an experiment described in a source paper and to obtain the same results—such a translation may be regarded as a practical one. Perhaps the translation is not couched in elegant terms; here and there several alternative meanings are given for a target word; a word or two may appear as a mere transliteration of original source words. Nevertheless, this translation has served its main purpose: a scholar in one land can follow the work of his colleague in another.

This limited scope has been set for us by our own as well as the machine's deficiencies. The heartbreaking problem which we face in mechanical translation is how to use the machine's considerable speed to overcome its lack of human cognizance. We do not yet really understand how the human mind associates ideas at its immense rate of speed; for example, how does it differentiate seemingly instantaneously between the two meanings of *calculus* in the following sentences: (1) The surgeon removed the staghorn calculus from the patient's kidney, and (2) The professor announced a new course in advanced calculus. And yet, a scheme for discerning such differences is what we must impart to the machine. Even if there now existed a completely satisfactory method for machine translation, today's machines would not be adequate tools for its implementation. They lack automatic transformers of printed text into coded signals, and their external storage devices are not up to the mark.

Before coming to grips with the mechanical translation problem, we investigated the types of difficulties we might encounter. We found that they fall into ten groups; so far, we have been able to cope—more or less successfully—with only the first five, which depend mainly on syntactic analysis. Some thought has been given to the far more difficult points

involving semantic considerations, but the short time spent in this area has not allowed us to transform the mathematical "existence solutions" into practical machine application. Thus, discussion of semantic problems is deferred. In this paper we are concerned mainly with syntactic analysis.

The Glossary

One of the indispensable accessories of MT is the construction of a specialized source-to-target glossary. The conventional publications would not suffice for MT, because their authors presuppose, on the part of the prospective user, (1) a wide acquaintance with the basic principles of the source language, (2) an excellent knowledge of the target language, and (3) a considerable familiarity with the terminologies—in both languages—relating to the special subject of the source text. These assumptions are hardly justified even in the case of the professional translator. It follows that a glossary, designed for use with an electronic processor, must embody an immense amount of information in addition to the material culled from the best existing dictionaries. But there is a limit to the amount of data that can be handled by even the most advanced type of electronic processor, if MT is to be at all expedient. It is imperative, therefore, that utmost care be used to select (1) the absolutely minimum quantity of information which would suffice for our needs, (2) the most economical (space and time-saving) form for representing it, and (3) the most suitable external media for its storage and retrieval.

Of far greater concern is the fact that we are not fully aware of the mental processes involved in the performance of the translation task. Yet a routine, paralleling these processes, must be prepared for insertion into the machine's memory. Unfortunately, the form of the glossary depends upon, and varies with, the particular translation scheme which is being developed. We would not venture to predict the date when our own glossary might assume its final—or even "passable"—shape. We are constrained, for the present, to use a small sample glossary, sufficient for trial runs on the computer. It is stored in the external memory and is arranged in groups, each of which lists the Satellites of a source Pseudo-root.[1] Each satellite is an entry corresponding to a source Stem which contains the pseudo-root in question. The temporary form, which each Glossary Entry has assumed so far, consists of the following items:

1. The Source Transform, which is a greatly contracted form of the original source stem.

2. Morphological information, designed to aid in the syntactical analysis of each sentence, as illustrated in Section B of Part II.

3. Predictions regarding future Occurrences. For instance, the Russian verb with stem служ is marked as frequently followed by an indirect object in the dative case and/or a complement in the instrumental; also sometimes by a verb in the infinitive.

4. One or more target correspondents (T) to the source stem.

(It is planned to expand this information to include diacritical material designed to aid in the semantic analysis of the sentence.)

PART I

Our program is being coded in two parts. Of these only the first, which consists of two sections, has been completed and tested.

Section A

The aim of this section is to investigate the nature of each Occurrence in a sentence and, for the case when the occurrence is a word, to perform a glossary look-up.

When an occurrence in a given Russian text is read into the machine—and we have reason to hope that this will be accomplished eventually by a fully automatic device—this source material is subjected to the following treatment within the computer.

1. An Identification Tag (t) is appended to the occurrence to indicate the page, sentence, and serial number. Its characters are counted and examined for indications anent its physical make-up. For instance, the machine examines whether the occurrence is a word, or perhaps, a punctuation mark, formula, etc. If a word, it notes whether it starts with a capital, or is an initial, whether it contains any indication of foreign origin. This orthographical material will be augmented and revised in succeeding steps to form General Specifications (GS). It is recorded in the internal memory space S_t allotted to the occurrence t.

2. If the current occurrence is not a word, this fact is indicated in the Profile Skeleton (PS) which will eventually be expanded to serve as a rough outline of the clause formation of the source sentence to which the occurrence belongs. If, moreover, the occurrence is identified as a period, a subroutine is consulted to determine whether this punctuation marks the end of the sentence. If such be the case, this fact is indicated

in the profile skeleton, and the sentence number is raised for storage in the succeeding tag numbers, t.

3. If the given occurrence is a word, a search is made in a Special List of frequently used words. If the word is found in the special list, the diacritical material accompanying it may show that it could be the leading word of one or more idioms. In that case, the requisite number of successive source occurrences will be compared to each of the indicated idioms; when agreement is found, the entire source idiom is replaced by the corresponding material and is thereafter treated as a single occurrence.

4. If the word is not found in the above list, it is decomposed into its pseudo-prefixes, pseudo-root (or roots), pseudo-suffixes, and source ending by means of corresponding lists stored in the internal memory (the pseudo-root and true source ending are determined by a rather complicated iterative scheme.)

The ending is replaced by the address β found alongside its listed counterpart. It is stored in S_t and will be used in part II.

Each pseudo-prefix and pseudo-suffix (if any) is replaced by a single character, consisting of 6 bits, and the combination of these characters (probably no more than 8) constitutes the transformation (Δ) of the original source word; y and z, the number of pseudo-prefixes and pseudo-suffixes, as well as Δ, are stored in S.

The remaining portion of the current word, constituting the pseudo-root, may have no characters at all. The glossary contains a group of satellites for a null pseudo-root, whose Extended Address, a_o, is used to represent it in the next step.

If the pseudo-root contains at least one character, it may not have been found in the list of pseudo-roots. In that case, the transliteration subroutine dictates the form of the correspondent to be stored in the normal position of the target T for the final printout. A suitable Signal of Peculiarity (δ) is stored in GS. The Correspondence Flag (c) in GS is set to zero.

If the pseudo-root has been located in the list, its counterpart is accompanied by an extended address, α, indicating where its group of satellites starts in the externally stored glossary.

5. The extended address, α, accompanied by the identification tag t, is intersorted with similar combinations, corresponding to the previously processed source words, in the Sorting File.

6. When all the internal space allotted for the sorting file is filled, a search is made throughout the entire glossary for the indicated entries. Since the time for such a transit throughout the glossary is formidable, and remains practically constant irrespective of the number of words to be looked up, it is obvious that an appreciable increase in internal storage space would result in a corresponding reduction in the look-up time per word. However, considering the high cost of internal storage devices, it might be more expedient to utilize inexpensive non-erasable external storage media with suitable buffering devices which allow for the simultaneous retrieval of information along several channels.

7. When the extended address α attached to t is reached during transit of the glossary, the routine searches for the entry corresponding to the $y.z.\Delta$ of the occurrence t. The correspondence flag c is set to 1 or 0 in GS, according to whether the search as been successful or not. In the latter case, the pertinent peculiarity signal is stored in GS and the tag t is placed in the normal position of the target T for final printout.

Illustration 1

As an example of the performance of this section of the program, we offer the text word расположение. Suppose this word occurs as the 7th word of the 4th sentence on page 1. The corresponding symbol for t is: 1.4.7. The occurrence is examined and found to be a word (not a punctuation mark etc.) composed of 12 letters. The Word Flag (w) in GS would be set to 1.

The machine determines that no such word appears in the special list of frequently used words. The occurrence is therefore examined for pseudo-prefixes. In this case, the combinations рас and по happen to be true prefixes. By referring to the stored list of pseudo-prefixes, the routine would replace рас by the letter V and по by the letter R. Unable to discover more prefixes, the routine would isolate the ending ие. Suppose that the list of endings indicates that information on this ending is stored in internal memory beginning at address 357; the machine then sets $\beta = 357$. The routine would proceed to identify ен as a suffix and replace it by the letter K. Finding no more pseudo-suffixes, the routine would store in $S_{1.4.7}$ the numerals 2 and 1, to indicate the number of prefixes and suffixes y and z; these would be followed by the transform Δ, which is VRK. The machine would then enter the subroutine for identifying the pseudo-root. In the present case, no difficulties would be encountered, as лож would be located at once in the list of pseudo-roots. In actual practice, a number of complications may arise. The given word may contain a polyroot; or what we assumed to be an ending may actually be

part of the pseudo-root; or we may not be able to locate the root at all. The sub-routine takes note of all these possibilities.

The root лож is replaced by α which would be, say, 2.47.3097, if the first member in the group of this root's satellites has the position number 3097 in the 47th block on the 2nd tape. To α we attach the tag t and intersort the result with the other contents of the sorting file. The entry in the internal memory, corresponding to the occurrence расположение, now has the two forms:

Storage	GS	β	$y.z$	Δ
$S_{1.4.7}$	Orthographic description	357	2.1	VRK
	α			t
Sorting File	2.47.3097			1.4.7

After a specified number of successive occurrences have been analyzed in this way, a transit will be made through the glossary. When the position 3097 of the 47th block on the 2nd tape is reached, the machine will locate and extract all the material corresponding to 2.1.VRK, i.e., all the information pertinent to the stem расположен. In GS, the correspondence flag c would be set to 1 to indicate that the search had been successful.

Section B

In this section we examine each word-occurrence of a sentence with two aims in view:

1. To assign to it all possible grammatical interpretations, which we call Temporary Choices, TC_j. These are arranged roughly in order of most probable appearance; j indicates the serial number. Information common to all TC_j is labeled with $j = 0$.

2. To indicate its significance in the profile skeleton. To accomplish the first aim we distinguish three types of words:

a. If a source word is found in the special list of frequently used words, its various TC_j are explicitly listed there.

b. For a word whose transform is found in the glossary, the TC_j are obtained by finding the common intersection between the possibilities given by its ending in the Table of Endings and those given by the morphological information of the stem's glossary material.

c. When a source word is represented merely by its transliteration, the TC_j must be made on the basis of its ending (and, possibly, its suffixes) only.

As regards the second aim, the TC_j which accompany a current word may reveal that it could be a possible indicator of a main clause, or subordinate clause, or a phrase. If such is the case, an appropriate signal is added to the profile skeleton, in which the nature of the non-word occurrences has previously been stored. The profile skeleton will be subjected to a crude analysis in Section A of Part II.

Illustration 2

Let us use again the word расположение, belonging under the heading 2b above. The glossary's morphological information indicates that its stem, расположен, could represent either:

1. An inanimate neuter noun, belonging to a declension class which is identified by the ending ие in the nominative singular; or

2. An adjective, of verbal origin, belonging to a declension class which is identified by the ending ый in the masculine nominative singular.

This material, used in conjunction with the information listed for the ending ие, leads the machine to eliminate the second possibility given by the glossary and to list the following two temporary choices:

TC_0	Noun, inanimate, neuter (common to both)
TC_1	nominative, singular
TC_2	accusative, singular

This word does not call for the insertion of a signal into the profile skeleton (PS).

PART II

Part II of the projected scheme, now in process of being programmed, has the purpose of analyzing the syntactical structure of each source sentence and of constructing a corresponding target sentence. While Part I works on at least several hundred source words in one pass—the number of such words is determined by the internal memory capacity of the machine—Part II, which is made up of three sections, works on one sentence at a time.

Section A determines, as far as possible at this stage, the clausal and phrasal structure within the sentence. Section B is an iteration scheme for examining syntactical relations among the Strings of a sentence. It processes each string in turn from the beginning to the end of each sentence, repeats this process if necessary and decides whether a translation has been effected. Thereafter Section C takes over, composes a target sentence and prints it out.

Types of Difficulties

We shall list, in order of increasing complexity, the ten difficulties which obstruct our path toward such a goal:

1. The stem of a source word is not listed in our glossary. This will occur quite often in our translation scheme, as we intend to omit from the glossary the majority of non-Slavic stems.

2. The target sentence requires the insertion of key English words, which are not needed for grammatical completeness of the source sentence. For instance, the complete Russian sentence: он бедный (literally *He poor*) should be translated as *He* (*is*) (*a*) *poor* (*man*).

3. The source sentence contains well-known idiomatic expressions.

4. The occurrences of a source sentence do not appear in the conventional order. Sober writing, without color or emphasis, employs few inversions. Our method, which consists of predicting each occurrence on the basis of the preceding ones, works quite well in that case. But such orderliness cannot be expected to hold for long stretches of the text.

5. The source sentence contains more than one clause.

6. Corresponding to an occurrence in the source sentence, more than one target word is listed in the glossary. Polysemy is, of course, recognized as a most formidable obstacle to faithful translation, whether human or mechanical. Hilarious (or heartbreaking, depending on your point of view) "malaprops" can be cited by the score to uphold the conviction of many linguists that the MT task is a hopeless one. Our faith in the inventiveness of the human brain makes us reject such gloomy forebodings.

7. The source sentence is grammatically incomplete. Such a situation is frequently the result of carrying on the thought from one or more previous sentences. To succeed, any MT scheme will have to be able to transcend the boundaries of a sentence (or a paragraph, or a section).

8. The source sentence contains ambiguous symbols. Since we are planning to confine our efforts to mathematical texts, such occurrences will be legion.

9. The syntactic integration of the source sentence results in an ambiguity. It is often of a type that could be resolved by semantic considerations; but sometimes it is inherent and thus not removable by any process.

10. A combination of difficulties is listed in this category. They are quite annoying but fortunately rare: misprints; grammatical errors; localisms; peculiar nuances; comments based upon the sound (or the spelling) of source occurrences, such as puns whose sense it is impossible to render into the target language.

We have thus grouped Russian sentences into 2^{10}, i.e., 1024 types. A sentence possessing none of the ten difficulties would be represented by type number $00000\ 00000_2$ whereas—at the other end—a sentence exhibiting all the difficulties would belong to type $11111\ 11111_2 = 1023_{10}$.

Our scheme is able to cope successfully—we believe—with the first five types of difficulties, which involve monosemantic occurrences, or at most idiomatic expressions. We can thus handle 32 types of sentences ranging in type number from $00000\ 00000_2$ to $00000\ 11111_2$.

Section A

In both sections of part I we kept up, for each source sentence, a profile skeleton which consists of a set of signals denoting to which special class (if any) each occurrence belongs. This tentative outline serves to indicate where the clauses and phrases of the sentence might have their inception. The routine in the present section carries out an iterative process which aims to set rough limits to these ranges, based upon the position in the sentence of its (1) punctuation marks, (2) conjunctions (3) actual, or possible, starters of main clauses, (4) actual, or possible, starters of subordinate clauses, (5) actual, or possible, predicates for each clause, and (6) actual, or possible, phrase starters.

As a result of this iterative scheme, the profile skeleton PS is replaced by a Temporary Profile (TP), in which each occurrence is associated with four designators:

1. Its clause number (C),

2. A Status Flag (*v*) to indicate whether the predicate of the clause has or has not occurred,

3. Its phrase number (P), and

4. A Backward Flag (*b*) to indicate a particular manner in which the string is to be handled during the process of syntactic integration.

In the event that the routine does not succeed in determining a clause or phrase number, it will insert a Signal of Uncertainty (X), which the routine in Section B will attempt to resolve.

Section B

At the conclusion of the preceding section, each source occurrence has been replaced by a string of information which will expand as we progress in the integration scheme. The string, at this point, contains several sets of data:

1. A set of general specifications, GS, consisting of

a. a word flag, *w*, indicating whether the occurrence was or was not a Word-utterance (W).

b. a correspondence flag, *c*, indicating whether or not the occurrence (or its transform) was located in the storage.

c. a peculiarity signal, δ, pointing out any significant feature of the occurrence.

2. A set of four designators, belonging to the temporary profile, TP.

3. If the occurrence was a W, its string will have in addition

a. a set of temporary choices, TC_j, giving all possible grammar interpretations of the source word.

b. a set of target correspondents, T, if the word (or its transform) has been located in the memory; otherwise the correspondent will be either

1) the transliteration of all (or part) of the word-utterance, if its pseudo-root is not listed; or else

2) the identification *t*, if its transform is not in the glossary.

c. a set of Glossary Predictions (GP), retrieved from the memory if such exist, each consisting of

1) a Grammar Essential (GE), indicating the predicted type of agreement with a temporary choice.

2) a Signal of Urgency (*u*), indicating the probability of fulfillment.

3) In many cases, a Pretarget Insert (PI), indicating—in coded form—the English word(s) which is (are) to precede the target(s).

In addition to the above items, there may be available at any stage of the iterative process the following information, which has been generated during the preceding portion of section B.

1. Foresight Predictions (FP). Expectations for future strings, based on past occurrences; e.g. a direct object is governed by a transitive verb. A foresight prediction contains at least three specifications:

a. Serial number, *k*, to distinguish the different foresights generated by the same string.

b. Urgency Code (U), designating the degree of necessity—or the proximity—of the expected string, (e.g. a code of 1 indicates: next occurrence or not at all).

c. Sentence Element (SE), such as Subject, Predicate, Complement, etc.

In addition to the above items, which are always present, a foresight prediction may contain data, in the form of

d. Morphological Specifications (MS) regarding animation, gender, number, etc.

e. An Insert Flag (*e*) to indicate whether or not an English preposition is to be inserted before the target correspondent, T.

2. Hindsight (H_1) regarding troublesome strings. When a Predictable Choice does not agree with any of the previous FP, Hindsight Entries about this Unexpected Choice are stored together with a Chain Flag (*f*) in H_1, to be considered with subsequent strings. Such apparent inconsistencies must all be resolved at the conclusion of the sentence, as a necessary (but not sufficient) criterion of successful syntactical integration. Here, too, are stored queries about strings whose syntax is questionable, even though they seemingly fulfill previous predictions. Entries in H_1 concerning these Doubtful Choices are not flagged.

3. Hindsight (H_2) regarding predicted alternate temporary choices. It may happen that more than one of the temporary choices TC_j agree with previously made predictions. In this case, one is selected as a link in the sentence structure and the others are stored for future consideration in the current (and subsequent) iterations.

4. Hindsight (H_3) regarding the remaining unpredicted temporary choices TC_j. These are "pigeonholed" for possible use in subsequent iterations.

5. Chain number (L). Whenever the machine, in proceeding through a sentence, encounters a string which it is unable to link with any previous predictions, it starts a new chain. There exist, however, five types of Unpredictable Choices which do not cause a new chain to be started. They represent (a) punctuation marks, (b) conjunctions, (c) adverbs, (d) particles, and (e) prepositions.

The Routine of Section B begins with the following steps:

1. All the hindsight entries, left in storage from the previous sentence, are cleared out.

2. The chain number L is set to 1.

3. The following two predictions, for the main clause, are stored as foresights:

k.U.SE

1.7.Subject

2.7.Predicate

where *k* is the serial number within the string; U is the urgency code (7 indicates the highest); and SE is the sentence element of the prediction.

We now attempt to determine the syntactic sentence structure by observing the following routine for each string. (The letter *q* will indicate the current String number; Q will denote this running coordinate as it ranges from 1 to *q*; K and J will denote, respectively, the *k* and *j* within the string Q.)

1. The routine examines the unfulfilled FP_{QK} within the current clause or phrase, in decreasing order of Q and increasing order of K. Each of them is tested for agreement with any of the T_j. The first TC which fits an FP is taken as the Selected Choice (SC) for this iteration. The successful FP is deleted. If there are several TC_j and none of them fit any FP_{QK} the hindsight information is examined for possible clues regarding the selection of a TC, to act as the SC. If no clue is found, TC_j becomes the SC. If, however, the string was marked by a backward flag *b*, the examination of foresight predictions is omitted. In this case the routine examines—in reverse order—the previous selected choices, SC, for agreement with TC_j. If the string is of the unpredictable type, TC_1 is taken as the SC.

2. The selected choice is indicated by QK_j, where Q is the number of the string where the successful prediction (if any) was made and K is the serial number of that prediction. If there is no such prediction for SC, both Q and K are designated as 0. The letter *j*, of course, represents the serial number of the chosen TC in the current string.

3. The chain number L is left unchanged, if the string has been predicted or is of the unpredictable type; otherwise L is raised by unity.

4. The designators C, *v*, and P of the temporary profile TP are revised—in the light of the SC—to form the Selected Profile (SP). The status flag *v* furnishes clues for the subsequent revision of the clause number C, and the syntactical integration determines the bounds of each phrase.

5. New predictions for the foresights are culled from three sources:

a. The temporary profile, TP, of the next string. If the TP indicates that a new clause is starting, the predictions of a new subject and predicate are entered as foresights.

b. The main routine. This may yield predictions of a general nature on the basis of the S. For example, if the SC is a noun, one such prediction states that the noun might be followed by a complement in the genitive case. If the SC is the subject, we examine whether the predicate has been found previously; if not, we add to the FP of the predicate the information that it must agree with the subject in person, number, gender, etc. Similarly, if the SC is the predicate, the FP of the subject—if unfulfilled—is amplified.

c. The glossary predictions, GP, accompany the chosen TC. Such predictions, if any, would arise from the peculiar nature of the original occurrence. For instance, a particular verb may govern the dative case.

6. The predictions yielded by a string are appraised against the entries previously placed in hindsight, in order to ascertain whether the former throw any light upon the difficulties and conflicts represented by the latter. If a partial explanation is obtained, a suitable notation is made alongside the corresponding entry. Whenever such an entry is completely explained away, it is deleted. If such a deletion takes place in H_1, chain number L is reduced by one, provided the entry bears the chain flag *f*. Sometimes, a rearrangement in order of the strings is indicated, as a result of the above appraisal.

7. The SC may indicate that a key target such as a noun or a verb, has not been explicitly stated in the source sentence. If such be the case, the routine determines the required Target Insert (TI) and constructs a corresponding New String. On the other hand, the SC may dictate the suppression of (a) target correspondent(s).

8. A target order number R is assigned to the string, to indicate the arrangement of occurrences in the target language. In general, the Rs are consecutive. If, however, the appraisal in step 6 calls for a rearrangement of strings, or if step 7 resulted in the insertion of a new string (or the suppression of an Old String) the affected Rs are renumbered in accordance with the desired sequence. Pretarget Inserts (PI), such as prepositions and articles, are not assigned an R. Their handling will be discussed in section C.

9. The TC_j which do not become the SC may, under certain circumstances, be disregarded. In the cases

where the routine directs the machine to retain them, they are entered into hindsight H_2 or H_3, according to whether they do or do not agree with any FP.

10. If the chain number L was raised in step 5, an appropriate query is entered into hindsight H_1 with a chain flag f. If the SC is a doubtful choice, suitable queries—unaccompanied by the chain flag—are also entered into H_1.

When the end of the sentence is reached, we need not embark upon another iteration if (1) the foresights do not contain unfulfilled predictions of urgency 6 and 7, and (2) the chain number is 1. (In that case H_1 should be clear of flagged entries.)

In this event, the selected choices for all strings are considered as Final Choices (FC) and the routine proceeds to section C. If however, another iteration is indicated, it investigates the H_2 information where resolution signals were placed during the previous iteration whenever some partial light was thrown upon any of its entries. As a result, one of the former selected choices is replaced by a more promising one, and the effect of that change is investigated. It is obvious that, if the number of unresolved entries in H_2 is high, it would be prohibitive to pursue all the possible combinations of selected choices. We therefore set a limit to the number of iterations we allow the machine to execute. In the unlikely event that all the possibilities inherent in the H_2 entries have been exhausted, the H_3 entries are attacked in the same manner.

Failure is conceded when the number of iterations already performed has reached the limit we had set for ourselves, or when the current set of selected choices repeats any of the previous sets (which are stored in the internal memory). In that case, the routine records a failure signal and indications of the types of errors encountered, to be printed out at the conclusion of Section C.

Section C

This section is devoted to the construction and printing of the target sentence.

1. The target correspondents listed with the final choices are arranged in the sequence given by R.

2. A subroutine supplies new pretarget inserts PI, in addition to those supplied by the foresights. These may be either English articles or prepositions. The set of PI (if any) are inserted in front of the proper correspondent for eventual printout.

3. A second subroutine affixes Pidgin Endings (E) to target correspondents whenever needed. (To conserve precious internal space, we regard—for the present—all English targets as grammatically regular. Thus the plural of *foot* will appear as *foot-s*.)

4. A count is made of all unresolved hindsight entries.

5. The resulting information is printed out. All inserts, whether PI or TI, are printed in parentheses. Words for which there are no target correspondents are enclosed in brackets. They may appear as some combination of the following word-sections:

a. a translated initial prefix

b. a transliterated full or partial stem

c. a transliterated full or partial word.

If the iterative routine failed to satisfy our criteria, this fact would be indicated by the failure signal and by the notations of the error types encountered. On the other hand, the satisfaction of the criteria is no guarantee that the result is a faithful translation, unless all three hindsights are clear and all occurrences are monosemantic. Since such eventualities will be extremely rare, we shall regard the tallies for the hindsight entries and the multiplicity of the printed meanings as a measure of the "goodness of fit" of our version.

Illustration 3

The chart given on the next pages outlines the syntactic integration of a sentence possessing the five types of difficulty which our routine is able to handle with some degree of success. On the other hand, it contains a number of polysemantic words, of which only a few can be resolved at present. For the remaining polysemantic words, we are forced to print out all the meanings contained in our glossary.

The chart incorporates all of the steps entailed in carrying out the first (major) iteration cycle involving the entire sentence. The reader may need guidance as regards the temporal sequence of these steps; we shall, therefore, review this sequence from the start of the process on through the handling of the first string of the sentence. The notes following the chart are designed to clarify situations which do not come up in string 1. The two lists appended to this report will furnish all pertinent definitions. All terms mentioned therein are capitalized in the material which follows.

1. The portions headed "Part I" and "T" list the material obtained prior to the initiation of the syntactic process.

PART I

q	Source Occurce	ϕ	Gen. Spcfcs. (GS) w	c	δ	j	Temp. Chs. (TC_j) Morph. Spcfcs. (MS) Ps	A	D	ts	G	Y	Z	Glss. Preds. (GP) GE & u g	a	d	i	x	m	PS
1	По- каз- а- ть		1	1	Cap	1	Vb		if					7	1	1	1			vi

Heading	Explanation
Occurrence	This portion, although overwritten in Part II, is retained here for the convenience of the reader. It indicates, on separate lines, the 1) Pseudo-prefixes, 2) Pseudo-roots, 3) Pseudo-suffixes, and 4) Ending of a Word-utterance which is listed in the Glossary.
ϕ	This Flag is not part of the routine. It is introduced for the convenience of the reader to indicate a Word-utterance which is found in the Special List.
GS	The General Specifications of this Occurrence indicate that 1) it is a Word-utterance, 2) its correspondent (T) is an English word retrieved from storage, and 3) it starts with a capital letter. (Cf. the headings in the List of Symbols).
TC_j:j	The index j remains at 1; this indicates that the given Occurrence has given rise to only one Temporary Choice.
TC_j:MS	The Morphological Specifications indicate that the Occurrence represents a verb (Vb) in the infinitive mood (if).
GP	The Glossary Predictions accompany the Stem of the Occurrence in Storage. The numbers in the various columns are Urgency Signals (u). Each of these, with the exception of the number headed by x, will give rise to Foresight Predictions to be stored in String 2. Their connotation is explained in the List of Terms, and will be clarified at the time they will be utilized.
PS	The Profile Skeleton, although overwritten in Part II, is kept in the chart for the convenience of the reader. It serves as a basis for determining the boundaries of the clauses and phrases within the source sentence.

PART II
Section C
Translation

q	T
1	de- mon- str- ate

Heading	Explanation
Section C T	In the Glossary, English correspondents are stored in compact form somewhat similar to the Source Transform. The decomposition is shown here.

2. The portion headed "TP" lists the components of the Temporary Profile, resulting from the iterative process performed on the data in the Profile Skeleton (PS) involving the entire sentence.

PART II
Sec. A.
TP

q	C	v	P
1	1	0	0

Heading	Explanation
C	The number 1 indicates that the given Occurrence forms part of the first clause.
v	This flag remains at the zero level until the Prediction of a predicate is fulfilled.
P	The zero indicates that the given Occurrence is not part of a phrase.

3. The following preliminary steps are executed by the routine prior to entering the (minor) iterative cycle involving the first String.

a. The Chain number (L) is set to 1.
b. The Hindsight is cleared of all content.
c. The Foresight Predictions (FP_{qk}) stored within String q are made by String q-1. When q is equal to 1, $FP_{1,1}$ and $FP_{1,2}$ are made by the routine in response to signals indicating the start of the first clause. They predict with utmost assurance (shown by U = 7) that a Subject and Predicate will occur within the clause.
d. The Selected Profile (SP) is temporarily set equal to the Temporary Profile.

4. The portion headed "Section B" indicates how the routine attempts to incorporate the given Occurrence as a link of a unifying Chain, representing the sentence structure.

PART II

Section B

		Fores't Prds. (FP_{qk})										Sel. Ch. (SC)				L	SP			Hindsight															
					Morph. Sps. (MS)							FP								H_1							H_2							H_3
																				Entry			Resol.				Entry					Res		
q	r	k	U	SE	V	A	G	Y	Z	e	Q	K	j	B		C	v	P	f	j	ES	h	Q	K	J	Q	K	j	ES	h	Q	j
1	1	1	7	Sbj							1	1	1		1	1	0	0	SE	Sbj	-	6	1										
	②	2	7	Pdc																	-	8	2										
																			1	Y	m	-	4	3									
																						-	10	3									
																						-	26	1									
2	24	1	7	Pdc	vf		N	3	s																								
	6	2	7	Cpl				a																									
	6	3	6	Cls				a																									
	24	4	3	Cpl				d		1																							
	9	5	3	Cpl				i		1																							

Heading	Explanation
SC:FP	The Q and K under this heading indicate that $FP_{1.1}$, calling for a Subject, has been responsible for the Selected Choice.
SC:*j*	The index 1 indicates that TC_1—a verb in the infinitive—fulfilling the above prediction, was selected.
L	The Chain number remains unchanged because the SC fulfilled a Foresight Prediction.
r	This signal is not part of the routine. It is introduced for the convenience of the reader to indicate the number of the String which caused the deletion of the Foresight Prediction alongside of which the *r* appears. The *r* accompanying $FP_{1.2}$ is encircled to indicate (to the reader only) that the prediction is amplified within String 2 in the light of the chosen Subject.
$FP_{2,k}$	Foresight Predictions, to be stored in String 2, are made on the basis of the Selected Choice in String 1. First, and most urgent, is the amplification of the Prediction of a predicate. There were no general grammar Predictions because in our routine a verb does not yield such. The remaining four Predictions came from the Glossary information, accompanying the Temporary Choice selected as SC for String 1. They predict that with utmost necessity a direct object (in the form of either a single word in the accusative case or of a clause) will occur and that a slight probability exists for the appearance of complements (i.e. indirect objects) in the dative and/or instrumental case, each followed by a suitable preposition (*e* Flag = 1). Since the *x* Flag in String 1 is unity, a suitable signal will be set to indicate that the next Positional Preposition will govern the locative case, unless the signal is turned off by an *x* equal to zero in a subsequent String.

Heading	Explanation
Hindsight Resol.	In general, the TC_j, of the Current String, and the newly made $FP_{q+1,k}$, stored in the next String, are examined to ascertain whether they throw any light on the perplexities, doubts, and conflicts recorded in the Hindsight columns. For $q = 1$, however, the Hindsight is empty at this stage of the process. (The resolutions indicated on the chart for String 1 will, of course, have been made by subsequent Strings.)
SP	In general, the doubts recorded in the selection of the components of the Profile are resolved in light of the SC and of the Hindsight resolution. For $q = 1$, this is seldom necessary.
Hindsight Entries	This information is derived by, and stored in, String *q* to be investigated by subsequent Strings.
H_1 Entries	The unflagged Entries in String 1 indicate that the SC was a Doubtful Choice. They are to be interpreted as follows: First Entry a. The *j* is blank to indicate that the Entry concerns the SC. b. The SE indicates that this part of the SC contains the doubtful item. c. The *Sbj* gives the specific Element of the SC which is doubtful. Second Entry a. The *j* indicates that TC_1 may be governed, not by $FP_{1,1}$ as chosen, but by some future FP. b. The Y indicates the part of some future FP which may govern TC_1. c. The *m* indicates that TC_1 may be the fulfillment of a Prediction for a verb in the infinitive mood.
H_2 Entries	The one Temporary Choice in String 1, necessarily taken as the SC, fulfilled (doubtfully) $F_{1.1}$ but not $F_{1.2}$. Therefore, H_2 has no Entry for String 1.
H_3 Entries	When a String has only one TC, there can be no unused Choices.

5. The portion headed "Section C" indicates the order and form in which the target correspondent will appear in the translated sentence.

	Section C			
	Translation			
q	R	PI	T	E
1	1	(To)	de- mon- str- ate	

Heading	Explanation
R	The order in which the correspondents will appear in the printed translation in general follows closely the order of the original Occurrences. On occasion, a particular Selected Choice (SC) and/or a Hindsight resolution may effect a deviation from the sequence. The routine in Section B indicates the proper order (R) of every correspondent in the printed sequence.
PI	A subroutine makes these crude Pretarget Inserts. They are printed in parentheses.
E	The Pidgin Endings assume all English words to be regular.

PART I

q	Source Occur'ce	ø	Gen. Spcfcs. (GS) w	c	δ	j	Temp. Chs.(TC j) Morph. Spcfcs. (MS) Ps	A	D	ts	G	Y	Z	Glss. Preds. (GP) GE & u g	a	d	i	x	m
1	По-каз-а-ть		1	1	Cop	1	Vb	if						7	1	1	1		
2	без	1	1	1		1	Pp							3					
3	по-с-ред-ств а		1	1		0 1 2 3	Nn	im			N	g n a	s p p					1	1
4	транс-ценд-ент-н-ого		1	0	Nsr	0 1 2	Aj	ps am			Ⓕ M	g a	s s						
5	анал-из-а		1	0	Nsr	0 1 2	Nn	am			M	g a	s s						
6	ос-нов-н-ые		1	1		0 1 2	Aj	ps im				n a	p p					1	
7	теор-ем-ы		1	0	Nsr	0 1 2 3	Nn	im			F	g n a	s p p						
8	ис-числ-ен-ия		1	1		0 1 2 3	Nn	im			N	g n a	s p p				1	1	
9	в-ер-о-ят-ност-ей		1	1	Cpr	0 1	Nn	im			F	g	p					1	1
10	и	1	1	1		1 2	Cj cd Av cd												
11	вместе с тем	1	1	1	Idm	1	Av cd												
12	глав-н-ые		1	1		0 1 2	Aj	ps im				n a	p p						
13	при-лож-ен-ия		1	1		0 1 2 3	Nn	im			N	g n a	s p p					1	
14	их	1	1	1	Spf	1	Pn	pn 3				g	p						

PART II

Sec. A	Section B																															Section C			
FTP		Foresʼt Prds. (FP_{qk})										Sel. Ch. (SC)				L	SP			Hindsight															
					Morph. Sps. (MS)							FP								H_1						H_2							H_3				
																				Entry				Resol.			Entry					Res.				
v P b	r	k	U	SE	V	A	G	Y	Z	e	Q	K	j	B		C	v	P	f	j	ES	h	Q	K	J	Q	K	j	ES	h	Q	j	R	Pl	T	E
v 0 0	1 2	1 2	7 7	Sbj Pdc							1	1	1		1	1	0	0	SE Sbj − 6 − 8 1 Y m − 4 − 10 − 26					1 2 3 3 1										1	(To)	de- mon- str- ate	
p 0 1	24 6 6 24 9	1 2 3 4 5	7 7 6 3 3	Pdc Cpl Cls Cpl Cpl	vf		N	3 a a d i	s	1 1	0	0	1		1	1	0	1																2		with- out	
0 1	3	1	7	Cpl				g			3	1	1		1	1	0	1															3	(the)	means	
0 X	4 4 6	1 2 3	1 0 3	Cpl Mfr Cpl		im	N	g g m	s	1	4	1	1		1	1	0	1								4 2 1 SE Mst + 5 2 2 2 A im Y a − 6									4	(of)	[trans- tzend- ent]	-al
0 X	5 5	1 2	1 0	Mst Cpl			Ⓕ	g g	s	1	5	1	1		1	1	0	1																5		[anal- iz]	
0 X	6 6	1 2	1 0	Cpl Mfr			M	g	s	1	2	2	2		1	1	0	0																6	(the)	bas- ic	
0 X	7 7	1 2	1 0	Mst Cpl		im		a g	p	1	7	1	3		1	1	0	0																7		[teor- em]	-s
0 X	8 8	1 2	1 0	Cpl Mfr		im	F	g a	p	1	8	1	1		1	1	0	0														3	8	(of)	calcul- ation/ us	
0 X	9 9 25	1 2 3	1 0 3	Cpl Mfr Cpl		im	N	g g i	s	1	9	1	1		1	1	0	0																9	(of)	pro- b- abi- ity	-s
c 0 X	10 10 26	1 2 3	1 0 3	Cpl Mfr Cpl		im	F	g g m	p	1	0	0	1		1	1	0	0	Ps Cj + 12 2 Ps Av + 12															10		and	
c 0 X											0	0	1		1	1	0	0															11		at the same time	
0 X 1											2	2	2	6	1	1	0	0															12	(the)	main	
0 X	13 13	1 2	1 0	Mst Cpl		im		a g	p	1	13	1	3		1	1	0	0															13		ap- plic- ation	-s
0 X	14 14	1 2	1 0	Cpl Mfr		im	N	g a	p	1	14	1	1		1	1	0	0															14	(of)	them	

continued on next page

continued from the previous page

PART I

q	Source Occur'ce.	ø	Gen. Spcfcs. (GS) w	c	δ	j	Temp. Chs. (TC_j) Morph. Spcfcs. (MS) Ps	A	D	ts	G	Y	Z	Glss. Preds. (GP) GE & u g	a	d	i	x	m
15	,		0	0	,														
16	котор- ые		1	1		0 1 2	Aj im						rl n a	p p					
17	с- хуж- ат		1	1		1	Vb	id	pr				3 p		2	2	1		
18	о- пор- ою		1	1		0 1	Nn im				F i	s			2	1	1		
19	мног- им		1	1		0 1 2	Aj ps						Ⓕ i d	s p					
20	зн- ан- иям		1	1		0 1	Nn im				N d	p						1	1
21	,		0	0	,														
22	–		0	0	–														
23	вот	1	1	1		1	Av dm								1				
24			1	0	Sis	1	Vb cp	id	pr										
25	мысл- ь		1	1		0 1 2	Nn im				F n a	s s						1	1
26	,		0	0	,														
27	мне	1	1	1		0 1 2	Pn pn	1							d 1	s s			
28	пред- лож- ени- ая		1	1		0 1	Aj	pv	pr				F n	s	2	2	1	1	
29	господ- ин- ом		1	1		0 1	Nn am	tl			M i	s			1				
30	С.		1	0	Inl	1 2	Nn pw Np												
31	Г.		1	0	Inl	1 2	Nn pw Np												
32	С- трог- ан- ов- ым		1	0	Cap	1	Nn pw												
33	.		0	0	Esp														

PART II

Sec. A	Section B																																Section C			
TP	Forest Prds. (FP_{qk})										Sel. Ch. (SC)				L	SP			Hindsight														Translation			
					Morph. Sps. (MS)						FP								H_1						H_2						H_3					
																			Entry			Resol.			Entry				Res.							
c v P b	r	k	U	SE	V	A	G	Y	Z	e	Q	K	j	B		C	v	P	f	j	ES	h	Q	K	J	Q	K	j	ES	h	Q	j	R	Pl	T	E
2 0 0											0 0 0				1	2 0 0																15		,	
2 0 0 1	16 16	1 7 Sbj ② 7 Pdc									16 1 1 13				1	2 0 0			SE Sbj + 21 2 Y a + 17															16		which	
2 1 0	17	1 7 Pdc			vf		N	3	p		17 1 1				1	2 1 0																	17		serve	
2 1 0	19 18	1 5 Cpl 2 5 Cpl					d i			1 1	18 2 1				1	2 1 0																	18	(by/ as the)	aid	
2 1 0	19 19 19	1 1 Cpl 2 0 Mfr 3 5 Cpl 4 3 Cpl			im	F	g i d i		s	1 1 1	19 3 2				1	2 1 0											19 4 1 SE Mst + 20						19	(to/for)	co- p- ious	
2 1 0	20 20	1 1 Mst 2 0 Cpl					d g		p	1	20 1 1				1	2 1 0																	20		know- ledge	-s
1 0 0		1 3 Cpl					m				0 0 0				1	1 0 0																	21		,	
1 0 0											0 0 0				1	1 0 0																	22		–	
pl 1 0											0 0 1				1	1 1 0																	23		here	
1 1 0		1 3 Cpl					d			1	2 1 1				1	1 1 0																	24		(is)	
1 1 0	25 25	1 5 Cpl 2 4 Cpl					n i				25 1 1				1	1 1 0																2	25	(the)	idea	
1 1 1		1 3 Cpl					m				0 0 0				1	1 1 1																	26		,	
1 1 1											0 0 1				2	1 1 1			1 Y d – 29 1 + 33														27	(to/for)	me	
pl 1 1 1											25 1 1 25				2	1 1 1																	28		suggest- ed	
1 1 1	27 29	1 5 Cpl 2 5 Cpl 3 3 Cpl					d i m			1 1	29 2 1				2	1 1 1																	29	(by)	Mr.	
1 1 1	30 30	1 1 Cpl 2 0 Mfr 3 3 Cpl			pw		g M i s d			1 1	30 2 1				2	1 1 1																2	30		[S.]	
1 1 1	31 31	1 1 Mfr 2 0 Cpl			pw		M i s g			1	31 1 1				2	1 1 1																2	31		[G.]	
1 1 1	32 32	1 1 Mfr 2 0 Cpl			pw		M i s g			1	32 1 1				2	1 1 1																	32		[S- trog- an-ov]	
0 0 0		1 1 Cpl 2 0 Mfr					g M i s			1	0 0 0				1	0 0 0																	33		.	

Note no.	q	Headings	Explanation
1	2	TP	The preposition starts a new phrase, as indicated by P = 1.
2		$FP_{2,k}$	The first Prediction is $FP_{1,2}$ revised in the light of the chosen Subject. The number 3 in the Y column indicates the 3rd person. The remaining FP were culled from GP in String 1. The Flag *e* indicates that a crude PI will be inserted before a T which fulfills a predicted dative or instrumental Complement.
3		SC	The Q and K are zero, because a preposition is an Unpredictable Choice.
4	3	TC_0	The *j* is set to zero to indicate that the information will be common to all the TC_j listed in String 3.
5		GP	The Grammar Essential *g* is not used for nouns or adjectives, since the routine always makes the prediction of a Complement in the genitive for these parts of speech. In Russian, adjectives may act as nouns.
6		H_2, H_3	When the Complement of a preposition is fulfilled, the other TC_j are disregarded and not entered in Hindsight.
7	4	Occurrence	The prefix and first suffix are non-Slavic.
8		TC_1:G	The encircled F means Masculine or Neuter.
9		TP and SP	The routine resolves the Uncertainty Signal X, in TP, concerning the phrase number P.
10		FP	The first two Predictions are made by the routine anent the noun in the preceding String. They were deleted simultaneously because, whether fulfilled or not, FP's with Urgency 1 or 0 are not kept beyond one String. (Cf. Urgency code in List of Terms). The third FP throws some light on the second query in H_1 but does not resolve it, as indicated by the negative sign in *h*.
11		H_2	Two alternate choices are noted here: First conflict. a. The SE indicates the part of some future SC which may throw light on the conflict. b. The *Mst* indicates the fact that the conflict will be resolved, if a future SC will supply a Master for the current SC. Second conflict. First Entry a. The A indicates the part in a future SC which may throw light on the conflict. b. The *im* indicates the fact that the conflict will be resolved, if a future SC will supply an inanimate Master for the current SC. Second Entry. a. The Y indicates the part in a future SC which may throw light on the conflict. b. The *a* indicates the fact that the conflict may be resolved, if no future SC fulfills the expectation of a Complement in the accusative case.
12		E	This is the Pidgin Ending for all non-Slavic adjectives.
13	5	TC:G	The Pseudo-suffix *H3* with a non-Slavic root will be taken as denoting a masculine noun. Any required non-Slavic feminine or neuter noun with this Pseudo-suffix will be included in the Glossary.
14		FP	These Predictions are supplied by the routine anent the adjective in the preceding String.
15		SC and H_2 resol.	The Master prediction was satisfied, thus resolving the first Entry in H_2 (Cf. Note 11). Unfortunately the Occurrence was a non-Slavic word and not represented in our Glossary. There was no information about its animation, and therefore the second Entry can not be resolved.
16		H_2 Entry	The Entry of the form 5.2.1, to indicate that $FP_{5,2}$ is also satisfied by TC_1, is omitted for the following reason: When a Master prediction is satisfied, there is at present no way to resolve this ambiguity, since considerations of semantics and of context would be involved.
17	6	SP and $FP_{4,3}$	Since none of the Predictions made by the Strings in phrase 1 are fulfilled by the current String, whereas a Prediction in the main clause is, the phrase number P is set to zero again. The unfulfilled Prediction in Phrase 1, namely $FP_{4,3}$, which bears a low U, is deleted.
18		TC_1 and H_1 resol.	Since TC_1 could serve as Subject, the fact is noted in the resolution column of the first Entry in Hindsight H_1 (in String 1).
19		SC and H_2	The SC throws partial light on the conflict in the H_2 column. (Cf. Note 11.)
20	7	TC:G	The suffix *EM* with a non-Slavic root will be considered as belonging to feminine nouns. (Cf. Note 13.)

Note no.	q	Headings	Explanation
21		TC and H_1 resol.	Although TC_2 could serve as the Subject, no note is made of this fact in the resolution of the first Entry in H_1, because it would be the Master of TC_1 in the previous String, which has already been recorded. (Cf. Note 18.)
22	8	TC_2 and H_1 resol.	Partial light is thrown on the first Entry in H_1.
23		TC_3 and H_3	This choice is pigeonholed.
24		T	There is no way to resolve the polysemy at the present stage of our investigation.
25	9	$FP_{9,3}$	This prediction deletes $FP_{2,5}$, as explained in the List of Terms under the heading "Urgency Code".
26	10	TC and H_1	The iterative scheme in Section A of Part II established the fact that this String did not start a new clause (since the old clause still had the Status zero). The current String could therefore represent either a coordinative conjunction binding two related SC, or an adverb. Since both Choices are Unpredictable, they are recorded as Doubtful (unflagged Entries in H_1).
27	11	GS:δ	This String represents an Idiom, which constitutes a single Word-utterance.
28	12	*b*	The Backward Flag was placed by the Profile routine in Section A of Part II. It indicates that, instead of examining the FP's, the routine will scan the previous SC in reverse order to establish a new link in the structural Chain.
29		SC:B	The entry in this column indicates that the SC in String 6 is conjunctively related to the current String, and the true nature of String 10 is thus established, as indicated by the resolutions in H_1.
30	14	GS:δ	The source word might under certain conditions be treated like a prepositive possessive modifier rendered in English by "their".
31		TC:D	The number 3 indicates the person.
32	16	FP	These Predictions were caused by the change in the clause number C of the previous String.
33		*b*; B; H_1	Similar entries have been explained in Notes 28, 29, and for String 1.
34	23	TP:υ	The Profile routine indicated that the current String must be followed by a New String representing a copulative verb. The Status Flag υ was set to 1 in order to effect the proper insertion. (Cf. Note 35.)
35	24	GS:δ	This String was inserted because of the υ Flag in the previous String (Cf. Note 34.)
36	25	FP	These Predictions were made by the routine anent the copulative verb in the preceding String.
37	27	TC_2	The locative case cannot be considered, since it is governed only by a preposition.
38		L	The Chain is broken, as the routine cannot account for the dative. (The $FP_{24,1}$ cannot be considered, since the two Strings are not in the same phrase.)
39		H_1:*f*	The Entry is flagged, since the Choice is Unexpected.
40		R	The rearrangement in the target order will be explained in Note 47.
41	28	*b*; B; R	Cf. Notes 28, 29, 47.
42	29	$FP_{29,1}$	This prediction throws partial light on the H_1 Entry in String 27. The *r* is indicated in the light of Note 47.
43		PI	Since the noun is animate, only one connotation of the preposition is used.
44	30	GS:δ and TC:A	This part is used in obtaining the signal *pw* (proper word).
45	31	FP	The appearance of an initial calls strongly for another proper noun. Hence the inversion of the usual order of Predictions after a noun.
46	32	CS:δ and TC:A	The fact that the capitalized word does not start a sentence causes A to become *pw*.
47	33		Since no other explanation offered itself for the H_1 resolution in 27, the previous explanation is accepted, the Entry is deleted, and L is reduced by one. This fact also causes the rearrangement of the two neighboring R's. The T's for these two strings reflect the result of this rearrangement.

End of Iteration 1. Since no FP's with U of 6 or 7 remain and the chain number L is equal to 1, no other iteration is necessary. The translation is printed out as indicated in the last three columns of the chart, followed by the tally of unresolved Hindsight Entries.

Received March 16, 1959
Revised December 16, 1959

Appendix I: List of Terms

ADDRESS, extended (a). The locator of an item on an external storage medium. Its form depends on the machine used. On the IBM 704 it consists of (1) a number, (2) block number, (3) serial position of item in the block.

CHAIN. A group of consecutive SC's characterized by the same chain number (L).

CHOICE,

doubtful. A Selected Choice which entails the recording of an unflagged query in Hindsight H_1.

final (FC). The Selected Choice in the last iterative cycle.

predictable. Not belonging to class of Unpredictable Choices.

selected (SC). A Temporary Choice selected as a link in the Chain during a current iteration.

temporary (TC). A grammatical interpretation of a Source Occurrence.

unexpected. A Predictable Choice which does not agree with any of the Foresight Predictions (FP).

unpredictable. A TC containing one of the following parts of speech: (1) a conjunction (2) an adverb (3) a particle, and (4) a preposition; or else (5) it is a punctuation mark.

CODE, urgency. Cf. Urgency.

ELEMENT, sentence. Cf. Sentence.

ENDING,

source. The true inflectional ending of a word in the source text.

pidgin (E). Regular ending affixed to stem of target correspondent, regardless of correct usage.

ENTRY,

glossary. A complete set of Glossary items corresponding to source Stem.

hindsight. A record of an Unexpected String, a Doubtful Choice, or a surplus Temporary Choice (cf. Hindsight).

ESSENTIAL, grammar (GE). A grammatical form called for by a glossary Prediction (e.g. an accusative called for by a transitive verb). Each type of GE has a separate location in the string reserved for it. A GE is predicted by storing an Urgency Signal in this location.

EXTENDED Address. Cf. Address.

FILE, sorting. The internal space allotted for sorting the Extended Addresses (*a*).

FLAG. A binary digit (i.e. either a 0- or 1-bit).

Backward (*b*). A 1-bit alerting the machine to examine the foregoing Selected Choices in order to establish the linkage of the current String.

Chain (*f*). A 1-bit accompanying a Hindsight Entry in H_1 to record an unexpected Choice.

Correspondence (*c*). A bit indicating whether the String contains target correspondents or not.

Insert (*e*). A 1-bit accompanying a Foresight Prediction to indicate that a suitable English preposition is to be used as a Pretarget Insert (PI).

Locative (*x*). A 1-bit in Glossary Predictions to indicate that the locative case is, 1) if the String represents a preposition, one of the cases governed by the preposition; or 2) if the String does not represent a preposition, to be used after the next Positional Preposition encountered in the sentence.

Status (*v*). A bit indicating whether the predicate of the current clause has turned up or not.

Word (*w*). A bit indicating whether the String represents a word or not.

GENERAL Specifications. Cf. Specifications.

GLOSSARY. The externally stored source-to-target dictionary used by the MT group at NBS. It is stored in a greatly compacted form and contains diacritical material designed to aid in the syntactic—and to a small degree in the semantic—analysis of source sentences.

GRAMMAR Essential. Cf. Essential.

HINDSIGHT. Antimale space allotted for storing in

H_1, Entries concerning Unexpected Choices or Doubtful Choices.

H_2, Temporary Choices (TC), other than the Selected one, which fulfill Foresight Predictions (FP).

H_3, Temporary Choices (TC), other than the Selected one, which do not fulfill any of the Foresight Predictions (FP).

IDENTIFICATION Tag. Cf. Tag.

INSERT,

target (TI). A target correspondent incorporated in a New String.

pretarget (PI). A word inserted before a target correspondent.

LIST. Internally stored one-to-one correspondences, yielding for each of the

endings, an address b for each Source Ending, enabling the machine to find, subsequently, the corresponding morphological information in the Table of Endings.

pseudo-prefixes, a 6-bit character, a substitute for each Pseudo prefix.

pseudo-roots, an Extended Address leading to the location of the first externally stored Satellite of each Pseudo-root.

pseudo-suffixes, a 6-bit character, as a substitute for each pseudo-suffix.

symbols, a definition (to be found in Appendix II).

terms, a definition (to be found in Appendix I).

special words, dictionary information. The arguments in this List consist of conjunctions, prepositions, particles, idioms, abbreviations, some adverbs, and words with ambiguous endings.

MORPHOLOGICAL specifications. Cf. Specifications.

OCCURRENCE, source. A combination of one or more characters in a source text. It may represent (1) a Word-Utterance; (2) punctuation mark or a set of such; (3) a symbol or a set of such; (4) a diagram or a set of such; etc.

POSITIONAL preposition. Cf. Preposition.

PREDICTIONS,

foresight (FP), information concerning TC's which are expected to occur somewhere in the sentence under consideration. Such information is derived either from rules of grammar incorporated in the machine instructions or from Glossary Predictions, or from the Temporary Profile.

glossary (GP), partial information retrieved from the Glossary or Special List and stored as part of a String, indicating what kinds of TC's are expected to occur somewhere before or after the current SC in the same sentence. A GP is recorded by assigning an Urgency Signal to a Grammar Essential. One String may contain several GPs.

PREPOSITION,

positional. One of a set of Russian prepositions which govern either the accusative or locative case.

PROFILE. The sequence of sets of designations, incorporated in each String of a sentence, which may throw light upon the ranges of its clauses and phrases.

Selected (SP). The Temporary Profile, revised during an iterative cycle in the syntactic integration process.

Skeleton (PS). The initial stage of a Profile, which bears one signal for each Occurrence, indicating the latter's significance in determining the clauses and phrases of the current sentence.

Temporary (TP). A sequence of sets of four preliminary designators assigning a rough clause (C) and phrase (P) number, as well as a Status (v)—and possible Backward (b)—Flag, to each Occurrence.

PSEUDO-ROOT. That portion (if any) of a word-Occurrence remaining after Ending, and Pseudo-suffixes are stripped off.

PSEUDO-PREFIX. One of a set of combinations of source letters which are frequently found before the Source Ending of words in the source language.

PSEUDO-SUFFIX. One of a set of combinations of source letters which are frequently found before the Source Ending of words in the source language.

SATELLITE of a Pseudo-root. A Glossary Entry listing the Transform of a Source Stem which contains the Pseudo-root in question.

SENTENCE element (SE). One of the following ingredients of a sentence: Subject, Predicate, Complement, Modifier, Master, Clause and Phrase.

SIGNAL,

peculiarity (δ). An indicator of some peculiar nature of an Occurrence, e.g. that it is a capitalized word, it is an initial, its root is not listed, etc.

uncertainty (X). Used instead of a clause or phrase number in a Profile, when the determination of that number is not possible.

urgency (u). One of the numbers 0 to 7, indicating the probability of a Glossary Prediction (GP), used to form Urgency Code U in FP, according to the following relation between u and U: $0\sim1$; $1\sim3$; $2\sim5$; $3\sim1$.

A u signal, 4 units higher than the above, indicates an alternate prediction of a clause.

SORTING file. Cf. File.

SPECIFICATIONS,

entry (ES). Signals in H_1 and H_3 to specify the type of query.

general (GS). Designators in a String, consisting of a Word (w) and Correspondence (c) Flags, as well as a Signal of Peculiarity (δ).

morphological (MS). Designators in the TC_j, FP and H_1 of a String, which deal with the grammatical interpretation of the original Occurrence.

STEM. The portion of a word remaining after the ending is removed.

STRING. The information, replacing the original Occurrence, which is available to the routine during the process of syntactic integration.

New. A String which is inserted during the process of syntactical integration.

Old. A String which is available at the beginning of the syntactic process.

TABLE of Endings. A tabulation of the morphological possibilities of each Source Ending.

TAG, identification (t). A serial number attached to each Source Occurrence of a text sentence. It consists of (1) page number, (2) sentence number, and (3) Occurrence number.

TRANSFORM, source (Δ). A contraction representing the Stem of a Source Occurrence in the external memory.

URGENCY code (U). One of the numbers 0 to 7, connoting the probability of a Foresight Prediction (FP) as follows:

7, must occur sometime

2, very likely to occur sometime

3, may occur sometime. An FP bearing this U is erased by a subsequent FP identical to it

1, will be the next Choice or won't occur at all. It is erased after the next SC.

A code of 6, 4, 2, and 0, indicates the same degree of Urgency as 7, 5, 3 and 1 respectively. Moreover, the even-numbered codes denote an FP alternate to the last preceding odd numbered FP in the same String (e.g. successive Us of 5, 4, 2 indicate that the second and third predictions are alternates for the first, so that if one of the three occurs, all three could be erased). An FP with $U > 4$ is not erased until the end of the iterative cycle, unless it, or one of its alternates are satisfied. An FP with $U = 6$ or $U = 7$, left unsatisfied at the end of a cycle, calls for another iteration.

URGENCY signal. Cf. Signal

UTTERANCE, word. Cf. Word

Word-utterance (W). One word or a set of consecutive words—in complete or abbreviated form—used as an entity (e.g. an initial, compound word, an idiom, etc.).

Appendix II: List of Symbols

α	Extended Address
A	Heading in TC_j and FP
a	Accusative case
Aj	Adjective, a Ps
Am	Animate
Av	Adverb, a Ps
β	Address of argument in Table of Endings
B	Signal in SC to indicate the String where concatenation was established during backward examination
b	Backward Flag
C	For all Strings other than the first in a sentence, C is the clause number in the Profile. For the first string of a sentence (since all sentences start with clause number 1), we shall use this symbol as a code: 1, declarative sentence 2, interrogative sentence 3, exclamatory sentence, etc. This is possible because the Temporary Profile is obtained as a result of an iterative routine, and the nature of the sentence is known before Part II is undertaken.
c	Correspondence Flag
Cap	Word starts with capital, in δ
cd	Coordinate
Cj	Conjunction, a Ps
Cls	Clause, an SE
co	Clause opener, in PS
cp	Copulative
Cpl	Complement, an SE
Cpr	Compound root, in δ
cw	Coordinate word, in PS
Δ	Source Transform
δ	Signal of Peculiarity
D	Heading in C_j
d	Dative case

dm Demonstrative

E Pidgin Ending

e A 1-bit to indicate that an English preposition is to be used as a PI

ES Entry Specifications

es End of sentence, in PS

Esp End-of-sentence period, in PS

F Feminine gender

f Chain Flag

FC Final Choice

FP Foresight Prediction

G Gender: M, F, and N. An encircled gender indicates the *two other* genders

g Genitive case

GE Grammar Essential, such as *g*, *a*, *d*, *i*, *x*, *m*

GP Glossary Predictions

GS General Specifications

H_1 Storage of queries anent Unexpected and Doubtful Choices

H_2 Storage of alternate predicted TC

H_3 Storage of surplus unpredicted TC

h (for convenience of the reader only) Signal for type of resolution in Hindsight: minus for partial, plus for complete

i Instrumental case

id Indicative mood

Idm Idiom, in δ

if Infinitive mood

im Inanimate

Inl Initial (a capital letter followed by a period), in δ

J A number indicating the *j*th TC in String Q

j A serial number of TC within a String *q*

K A number indicating the *k*th FP in String Q

k A serial number of an FP within a String *q*

L Chain number, indicating the degree to which the syntactic integration lacks cohesion

l locative case

M masculine gender

m A Flag, indicating whether or not a verb in the infinitive is predicted

Mfr modifier, an SE

Mst Master, an SE

N Neuter gender

n Nominative case

Nn Noun, a PS

Np Not part of speech

Nsr Non-Slavic root, in δ

ϕ (for convenience of the reader only) A 1-bit to show that the W was found in the Special List

P Phrase number in a Profile

p Plural number

pc Postpositive copulative implied, in PS

Pdc Predicate, an SE

PI Pretarget Insert

Pn Pronoun, a PS

pn Personal

po Phrase opener, in PS

Pp Preposition, a Ps

Pr present tense

PS Profile Skeleton. A set of signals, in addition to punctuation marks, such as

- co clause opener
- cw coordinate word
- es end of sentence
- po phrase opener
- pc postpositive copulative implied
- vf verb finite
- vi verb infinitive

Ps Part of speech

ps Positive degree

pv Passive voice

pw Proper word

Q Running coordinate for *q*

q The serial number of the Strings

R Target order number, indicating sequence in which the English correspondents will be printed out

r (for convenience of the reader only), indicator of the String in the consideration of which a given FP is deleted. An encircled *r* indicates the String where the FP is revised.

rl	Relative
S_t	Storage space in internal memory slotted to information about Occurrence *t*
s	Singular number
Sbj	Subject, an SE
SC	Selected Choice
SE	Sentence Element, in an FP
Sis	String inserted by syntax, in δ
SP	Selected Profile
Spf	Special possessive form, in δ
T	Target correspondent
t	Identification Tag of Occurrence
TC	Temporary Choice
TI	Target Insert
tl	Title
TP	Temporary Profile
ts	Tense
U	Urgency Code, in FP
u	Urgency Signal, in TC
V	Heading in FP
v	Status Flag
Vb	Verb, a Ps
vi	Verb infinitive, in PS
vf	Verb finite, in PS and under heading V in FP
W	Word-utterance
w	Word Flag
X	Signal of Uncertainty, in PS
x	A 1-bit to indicate that the locative case is governed either at once, if the TC represents a preposition, or after the next Positional Preposition (which is cf in List of Terms).
Y	A heading in TC and FP
y	Count of Pseudo-prefixes in a word-Occurrence
Z	A heading indicating grammar plurality
z	Count of Pseudo-suffixes in word-Occurrence.

Notes

This work was sponsored by the Office of Ordnance Research, Department of the Army. The author acknowledges with deep gratitude the gracious and generous aid of her chiefs and colleagues, Drs. Edward W. Cannon, Franz L. Alt, Don Mittleman, and Henry Birnbaum who devoted an extraordinary amount of time and effort in writing large portions of this report and in painstakingly revising the rest. Special thanks are also due to her collaborators Mrs. Patricia Ruttenberg, who single-handedly coded Part I of the scheme described herein, to Dr. Leroy F. Meyers, who offered many valuable suggestions for improving the scheme, and to Mrs. Luba Ross for her amazingly patient and competent attention to details while preparing the manuscript for publication. Because of the long delay between completion of the manuscript and its appearance in print, this paper no longer represents the author's latest treatment of the problem.

Numbers in brackets after note numbers are those appearing in the original, uncut version of this paper. However, we print the original reference list in full for its documentary value; this means that the reference numbers in this abbreviated text are not always sequential.

1. The List of Terms and List of Symbols at the end of the paper may enable the reader to identify unfamiliar expressions. Technical words to be found therein are capitalized when first encountered in the text.

8
A Preliminary Approach to Japanese-English Automatic Translation

Susumu Kuno

1 Four Stages of Automatic Translation

The proposed procedure for automatic translation of a Japanese linear text into English can be divided into four stages: (1) automatic input editing, (2) automatic segmentation with morphological analysis, (3) syntactical analysis, and (4) transformation with output editing, including semantic transfer.

2 Forms of Input Texts

It is apparent that input texts which are in accord with a commonly accepted writing system are better in the sense that they need less pre-editing before they are fed into a machine. Because the standard vernacular writing system of Japanese makes no use of spaces between words and because *kanas* (syllabic Japanese characters) and *kanjis* (ideographic Chinese characters) are used instead of Roman letters, it is necessary to devise a method of automatically cutting into its components the unsegmented sentence, written in kanas and kanjis.

In the standard writing system, 71 kanas and 1850 kanjis are used. A kana is a syllabic grapheme, which broadly speaking, corresponds to a combination of a consonant and a vowel in speech. A kanji, on the other hand, is an ideogram with one or more pronunciations attached to it. The pronunciation of kanjis, however, is usually governed by their combination with other kanjis, or with declensional kana endings. Graphemes used to represent so-called grammatical forms are in most cases kanas, but every kanji can be replaced by one or more kanas which represent its pronunciation, so that one utterance can be represented by a variety of sentences ranging from those in which kanjis are used wherever possible to those in which no kanjis are used, including a number of intermediate possibilities. In order to prevent the size of the automatic dictionary from being too much enlarged by storing two or more representations for each morpheme (when both kana and kanji representations are allowable), it is necessary to regulate the form of input kana and kanji texts. In the proposed system of automatic translation, two possibilities are considered: (a) kana texts in which no kanjis are used; and (b) kana-kanji texts in which kanjis are used wherever possible according to the official directives about the use of kanas and kanjis.[1]

The manual process of reducing Japanese texts to one or another of the input forms (a) and (b) will be called "pre-editing," and will be distinguished from automatic input editing, which is the first stage of the procedure for automatic translation, in which kanas and kanjis are transformed to tokens which are accepted by a computer and are convenient for subsequent automatic segmentation.

3 Dictionary and Auxiliary Items

The automatic dictionary is expected to consist of dictionary items arranged in alphabetical order with various grammatical codes and English correspondents. *Dictionary items* are units of Japanese established for the purpose of automatic segmentation, roughly corresponding to what might be termed lexical as opposed to grammatical forms. A file of *auxiliary items* is to be stored apart from the dictionary file, consisting of units of Japanese corresponding roughly to so-called grammatical forms. These are to be arranged in small groups according to *distribution types*, also for the purpose of automatic segmentation.

All dictionary and auxiliary items are expected to be coded according to *distribution types*, *function types*, and *transformation types*. Distribution types are categories or items established on the basis of their combination with contiguous items. Function types are categories of words established on the basis of potential roles they may fulfill in sentences; i.e., on the basis of their prediction and fulfillment of syntactical relationships. Units of syntactical analysis are called *words*. They do not necessarily correspond to items, which are essentially units of automatic segmentation. A word consists of one or more items, and the function type of a word is a product of the function type

codes of its component items. For instance, if the dictionary item "hanas" (to speak), having a function type code for *verb*, is combined with the auxiliary item "u" (verbal final-attributive suffix), which has a function type code for *final*, the function type of the resulting word "hanasu" will be *final verb FT*. Although the distribution type of an item seems to be rather closely connected with its function type code, it is necessary to distinguish between the two. In order to avoid confusion in terminology, names of distribution types and function types will always end with *DT* (distribution type) and *FT* (function type) respectively; e.g., *substantive DT*, and *substantive FT*. "*DT substantives*" and "*FT substantives*" refer to members of the *substantive DT* and *substantive FT* categories. Transformation types are categories of words established on the basis of their roles in the structural transfer between two languages, pertaining mainly to word order, omission of words in the source language, and insertion of new words in the target language.

4 Automatic Input Editing

In the first stage of the automatic language translation process, each kana in an input text will be transformed into two tokens for two Roman letters so as to preserve a one-to-one correspondence between kanas and their correspondent Roman letters. In a kana-kanji input text, however, each kanji will be transformed into an irreducible unit token. For instance, three kanas (shown in entry 1, table 8.1) in a kana text will be replaced by tokens for "hanasu", which has six characters: "h", "a", "n", "a", "s", and "u". On the other hand, a kanji and a kana (entry 2, table 8.1) in a kana-kanji text will be replaced by tokens for "(hana)su", which has three characters: "(hana)", "s" and "u". The replacement of each kana by two reduced tokens in a kana and kana-kanji text is due to the assumption that the Japanese inflectional system is better analyzed on the level of Roman letters than on the level of irreducible kanas, with fewer varieties of suffixes and fewer rules of permissible combinations with canonical stems, and with fewer possibilities of homographic verbal stems. Replacement of each kanji by one irreducible token, on the other hand, is due to the expectation that in the prospective analysis no kanji will contain more than one "morpheme".

Table 8.1
Kana and Kanji Reference

Entry	Kanas and Kanjis
1	ハナス
2	話ス

The above mentioned transformation may be done automatically by means of a kana typewriter or a kana-kanji typewriter equipped with magnetic tape or other memory device for internal conversion to the desired representation. Input provisions will vary according to the type of computer used.

5 Automatic Segmentation

The second stage of the process pertains to the automatic segmentation of a continuous run of tokens for representations of kanjis and kanas. The proposed method of automatic segmentation is based on the prospect that in our analysis auxiliary items will be shorter in length and fewer in number than dictionary items, and that no problem will be caused by assuming that every "phrase" in a sentence begins with a dictionary item whose average length is greater than that of auxiliary ones, or by including "prefixes" in the category of dictionary items, as they are very scarce in Japanese.

A method is proposed providing for a "find the longest matching dictionary item" (subsequently referred to as "find the longest") operation combined with the testing of immediately following sequences of tokens against predicted auxiliary items (referred to as "predictive testing"). The distribution type of the longest matching item is examined, and then a string of tokens immediately following a "matched segment" (i.e., a segment corresponding to a dictionary item found by the previous operation) is tested to determine whether it is initiated by any of the auxiliary items which are predicted on the basis of that distribution type.

Suppose the distribution type of the longest matching item found at the beginning of a text is *substantive DT* which is assumed, as a simplified model, to allow nothing to follow except one or more *DT particles*. This item is first considered to be *relevant* on the basis that every dictionary item can be combined with a preceding space, actual or hypothetical. The next step is to go to a subroutine in which each *DT particle* predicted to succeed this distribution type is tested against a string of tokens immediately following the matched segment. If a segment or segments matched by one or more *DT particles* are found, they are sep-

arated, and it is assumed that each matching item, with the exception of the last *DT particle* found is *locally valid* on the basis that it is followed by an item whose combination with it is permissible in the language. Then comes a new "find the longest" operation, and the longest matching item, if found, is tested against the item last separated to determine whether or not the combination of the two is permissible. If it is permissible, the newly found item is said to be relevant, and the preceding item is said to be locally valid; if it is not, the newly found longest matching item is first considered to be *irrelevant* and the second longest matching item is then sought.

If, on the contrary, no segment matched by any of the *DT particles* is found following the *first DT substantive*, the dictionary item is tested as to whether it can be followed directly by another dictionary item with no intervening auxiliary item(s). If the answer is yes, a new "find the longest" operation is performed upon the remainder of the sentence, and the matching item found is used as a key for determining the local validity of the preceding item. If the answer is no, the first item (*DT substantive*) is considered to be *invalid* and the second longest matching item is sought.

When the whole sentence has been cut into segments successfully matched by locally valid items, it is assumed that these items are wholly *valid* on the basis of the proposed program of automatic segmentation. Both structurally and semantically valid items are said to be *correct*.

One or more matched segments in a sentence beginning with a dictionary item and ending with an auxiliary item (if any) immediately before the next dictionary item will be said to form a *joint* whose *nucleus* is a segment matched by a dictionary item, and whose *subsidiary* is a segment matched by an auxiliary item. Joints are classified according to the distribution type to which their nuclei belong. The longest possible combination of a nucleus and subsidiaries in each type of joint is called a *maximum joint*. Rules of the combination and ordering of auxiliary items as subsidiaries in maximum joints are studied in detail. Auxiliary items are classified according to their relative ordering in a joint; e.g., as to which elements must precede, or are prohibited from following others, etc.

5.1 Inclusion Marks

These are various ways of programming the procedure to be utilized in automatic segmentation. The technical problem will not be discussed in detail, but a brief outline of a proposed method follows.

Table 8.2
Illustration of Inclusion Marks

Address	Entry	Inclusion Mark		English Correspondent
a1	ha	0		tooth, leaf
⋮				
b1	hana	2,	a1	flower, nose
b2	hanabana	0		flowers
b3	hanabanasi	2,	b2	brilliant
b4	hanabi	0		fireworks
b5	hanakago	4	b1	flower basket
⋮				
c1	hanas	1	b1	to talk
c2	Hanasi	1	c1	story
c3	hanataba	0		bouquet
c4	hanawa	0		wreath
c5	hanaya	2,	b1	flower shop
c6	hanayaka	2,	c5	gay, splendid
c7	hanayome	0		bride

When a machine with a large addressable memory is available, it is comparatively easy to incorporate the "find the longest" instruction in a table look-up process. There may be cases in which the longest matching item found is not correct, because during the operation the proper item has been erroneously associated with a string of letters immediately following (which may or may not be meaningful). For example, in automatically segmenting "hanayaki", "hanaya" ("flower shop") will be selected as the longest matching item for an automatic dictionary which has entries such as those shown in table 8.2. But this segmentation is wrong because in the above text, a cut should most probably be made between "hana" ("flower") and "ya" (*particle DT* "*and*") and "ki" ("tree"), on the assumption that "hanaya" will never be followed by any item beginning with "ki". Likewise, "hanakago" as one unit means "flower basket", but it may also be "hana ka go", "hana" meaning "flower" and "ka" being a *DT particle* meaning "or". The final segment, "go," may either mean "five" or the game of "go", or may constitute a nonsensical sequence of letters detached from the beginning of an item with a form such as "goma" ("sesame"), "gomi" ("dust"), or "gobo-u" ("burdock").

To prevent an erroneous segmentation of this kind, what may be called an "inclusion mark" may be prepared for every dictionary and auxiliary item. This mark consists of a single digit indicating where an

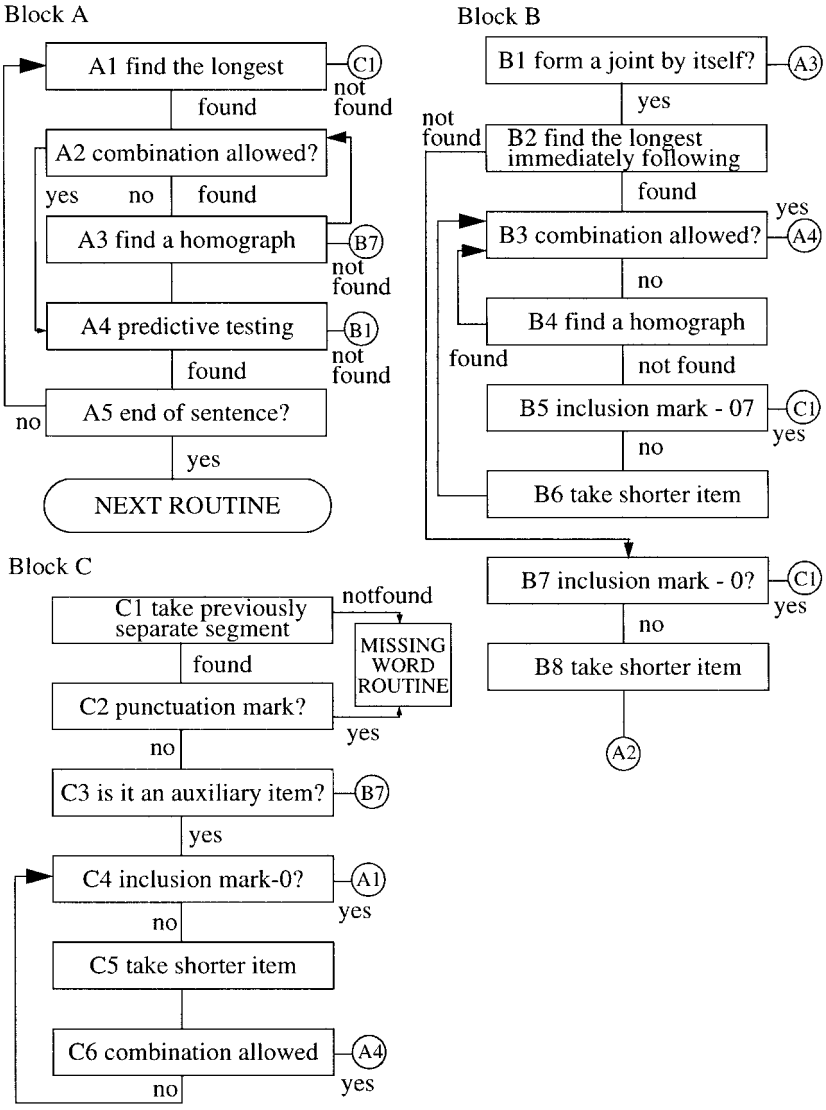


Figure 8.1
Flowchart of automatic segmentation.

alternative cut may be made (counting backwards in the sequence of letters), and showing the dictionary location of the shorter item thence produced (see table 8.2). A cut is made only when both of the resultant two segments are legitimate. "Hanawa", for instance, has an inclusion mark "o" on the assumption that "hana" will never be followed by any item beginning with "wa".

5.2 Program of Automatic Segmentation

Figure 8.1 contains a flow chart for finding the longest matching item in the dictionary and for predictive testing of auxiliary items, with previously made segmentation corrected on the basis of inclusion marks. The program is divided into three blocks. Block A is for normal repetition of cycles for "finding the longest" and succeeding "predictive testing"; block B is for rejecting the longest matching item and taking a shorter one as indicated by the inclusion mark; block C is for correcting retrospectively segmentation previously made. An explanation of the notations used in the flowchart is given below.

A. Steps in the operation are numbered A1, A2, ...; B1, B2, ...; etc., the initial Roman letters indicating block numbers.

B. "Combination allowed?" (A2, B3, C6) means "Is the combination of a newly found item with a previously found item allowed?" If the answer is yes, the newly found item is considered to be relevant, and the preceding item to be locally valid. If the answer is no, the newly found item is considered to be irrelevant. Items found through "predictive testing" (A4) do not undergo this operation since it is clear that they are relevant.

C. Punctuation marks and spaces are assumed to be included in auxiliary items.

D. Segments in the text which are matched by dictionary and auxiliary items found by the "find the longest" (A1 and B2) and "predictive testing" (A4) operations respectively are to be separated and stored in temporary storage in the order of their text occurrence.

E. "Form a joint by itself?" (B1) means "Can the item form a joint by itself, without being followed by any auxiliary item?"

F. "Find the longest immediately following" (B2) means "Find the longest matching item at the head of the remaining text."

G. "Take shorter item" (B6, B8, C5) means "Bring into register, and place in temporary storage the shorter item as indicated by the inclusion mark of the item found through operation of steps B2, A1, and C1 respectively, and modify the remaining text".

H. "Take previously separated segment" (C1) means "Bring into register a foregoing previously separated segment, taking it out of temporary storage and at the same time returning it to the remaining text."

I. If "end of sentence" (A5) is reached, all the previously found items are considered to be wholly valid.

J. The missing word routine has not yet been studied.

Figure 8.2 shows the process of automatic segmentation of "*sorehahana-yaka'u'e kiyani'aru*". The "step numbers" in the first column correspond to the step numbers in the flowchart of Figure 8.1. Matched segments and tested inclusion marks are shown in the fourth column under the heading "register". Matched segments are separated and stored in "temporary storage" (fifth column). If nothing is stored in the "register", "temporary storage", or "remaining text", a "D" is used.

Correct segmentation of the input sentence will yield "sore" (*substantive DT* "it")/"ha" (thematic *particle DT*)/"hanaya" (*substantive DT* "flower shop")/"ka" (*particle DT* "or")/"'u'ekiya" (*substantive DT* "gardener's")/"ni" (*particle DT* "at")/"'ar" (*consonantal verbal stem DT* "to be")/"u" (*consonantal verbal suffix DT*)/"." (*period DT*). Difficulty may be expected because it is probable that stored in the dictionary are both an item longer than "hanaya", that is "hanayaka" (*adjectoverb DT* "gay, splendid"), having the inclusion mark "2" and the address of "hanaya" (see table 8.2), and an item longer than "'ar" (*joint former DT* "a certain, unnamed") that belongs to the distribution type which can form a joint by itself, but which never stands before a period.

In the process of automatic segmentation of this sentence, "hanayaka" is considered to be relevant because its combination with the preceding item "ha" is permissible. It is then found that "hanayaka" is not followed by any predicted auxiliary items, that it cannot form a joint by itself, and that it has no homographs. Following the "inclusion mark-0?" and "take shorter item" operations, the shorter item "hanaya" is chosen, and the automatic segmentation continues correctly.

The input sentence in figure 8.3 is "*hanayahaga'utukusi'i*". Correct segmentation will be "hana" (*substantive DT* "nose, flower")/"ya" (*particle DT* "and")/"ha" (*substantive DT* "tooth, leaf")/"ga" (*subjective particle DT*)/"'utukusi" (*Adjectival stem DT* "beautiful")/"'*i*" (*adjectival suffix DT*, present final-attributive)/"." (*period DT*). Difficulty may be caused if an item is stored in the dictionary which is longer than "hana"; that is "hanaya" (see table 8.2), and if the local validity of "hanaya" is assumed by a following "ha" (thematic *particle DT*), and it may also be caused if the next "find the longest" operation yields "ga" (*substantive DT* "month"), on the basis of which the preceding "ha" is assumed to be locally valid. It is at this point of segmentation, when "hanaya ha ga" is reached, that it comes to a dead end, which makes possible correction of the previous made segmentation. "Ga" (*substantive DT*), however, is not followed by any predicted auxiliary items, and cannot form a joint by itself. Since "ga" is not found to be locally valid, the routine for correcting the previously made segmentation "hanaya ha ga" is initiated.

In each example, the entire sentence has been cut into wholly valid segments, which match exactly correct segments of the sentence.

There would be less difficulty in automatic segmentation if the above examples were kana-kanji texts in which tokens for the two kanjis and the kana, "(hana) (ya) ka" ("flower shop" and "or") (see entry 1, table 8.3), were different from those for the four

step number	operation	fulfilled condition	register	temporary storage	remaining text
A1	find the longest	FOUND	sore	sore	hahanayaka'u'ekiyani'aru.
A2	combination allowed?	YES		"	"
A4	predictive testing	FOUND	ha	sore ha	hanayaka'u'ekiyani'aru.
A5	end of sentence?	NO		"	"
A1	find the longest	FOUND	hanayaka	sore ha hanayaka	'u'ekiyani'aru.
A2	combination allowed?	YES		"	"
A4	predictive testing	NOT FOUND	Δ	"	"
B1	form a joint by itself?	NO		"	"
A3	find a homograph	NOT FOUND	Δ	"	"
B7	inclusion mark-0?	NO	2	"	"
B8	take shorter item		hanaya	sore ha hanaya	ka'u'ekiyani'aru.
A2	combination allowed	YES		"	"
A4	predictive testing	FOUND	ka	sore ha hanaya ka	'u'ekiyani'aru.
A5	end of sentence?	NO		"	"
A1	find the longest	FOUND	'u'ekiya	sore ha hanaya ka 'u'ekiya	ni'aru.
A2	combination allowed?	YES		"	"
A4	predictive testing	FOUND	ni	sore ha hanaya ka 'u'ekiya ni	'aru.
A5	end of sentence?	NO		"	"
A1	find the longest	FOUND	'aru	sore ha hanaya ka 'u'ekiya ni 'aru	
A2	combination allowed?	YES		"	"
A4	predictive testing	NOT FOUND	Δ	"	"
B1	form a joint by itself?	YES		"	"
B2	find the longest, immediately following	NOT FOUND	Δ	"	"
B7	inclusion mark-0?	NO	1	"	"
B8	take shorter item		'ar	sore ha hanaya ka 'u'ekiya ni 'ar	u.
A2	combination allowed?	YES		"	"
A4	predictive testing	FOUND	u	sore ha hanaya ka 'u'ekiya ni 'aru	.
		FOUND	.	sore ha hanaya ka 'u'ekiya ni 'aru .	Δ
A5	end of sentence? ⇓ NEXT ROUTINE	YES			

Figure 8.2
Example 1 (automatic segmentation).

kanas (entry 2), "hanayka" ("gay, splendid"); tokens for the kanji and kana (entry 3), "('a) ru" ("am, is, are"), from those for the two kanas of entry 4, "'aru" ("a certain, unnamed"); and tokens for the kanji and kana of entry 5, "(hana)ya" ("flower" and "and"); from those for the two kanjis, "(hana) (ya)" ("flower shop"), (entry 6). We no longer have an option for arriving at an item longer than the correct one. Generally speaking, automatic segmentation of kana-kanji texts seems to be far easier than that of kana texts.

6 Syntactical Analysis

The method of syntactical analysis proposed is that of predictive analysis, originally conceived by Rhodes,[2] adopted and developed at Harvard University for Russian by Sherry[3] in collaboration with Oettinger, for English by Bossert, Giuliano and Grant,[4] with theoretical implications of the method having been investigated by Oettinger.[5,6]

One peculiarity of predictive analysis as applied to Japanese is that it seems more convenient to start

step number	operation	fulfilled condition	register	temporary storage	remaining text
A1	find the longest	FOUND	hanaya	hanaya	haga'utukusi'i.
A2	combination allowed?	YES		"	"
A4	predictive testing	FOUND	ha	hanaya ha	ga'utukusi'i.
A5	end of sentence?	NO			"
A1	find the longest	FOUND	ga	hanaya ha ga	'utukusi'i.
A2	combination allowed?	YES		"	"
A4	predictive testing	NOT FOUND	Δ	"	"
B1	form a joint by itself?	NO		"	"
A3	find a homograph	NOT FOUND	Δ	"	"
B7	inclusion mark-0?	YES	0	"	"
C1	take previously sep. seg.	FOUND	ga	hanaya ha	ga'utukusi'i.
C2	punctuation mark?	NO		"	"
C3	is it an auxiliary item?	NO		"	"
B7	inclusion mark-0?	YES	0	"	"
C1	take previously sep. seg.	FOUND	ha	hanaya	haga'utukusi'i.
C2	punctuation mark?	NO		"	"
C3	is it an auxiliary item?	YES		"	"
C4	inclusion mark-0?	YES	0	"	"
A1	find the longest	FOUND	ha	hanaya ha	ga'utukusi'i.
A2	combination allowed?	NO		"	"
A3	find a homograph	NOT FOUND	Δ	"	"
B7	inclusion mark-0?	YES	0	"	"
C1	take previously sep. seg.	FOUND	ha	hanaya	haga'utukusi'i.
C2	punctuation mark?	NO		"	"
C3	is it an auxiliary item?	NO		"	"
B7	inclusion mark-0?	YES	0	"	"
C1	take previously sep. seg.	FOUND	hanaya	Δ	hanayahaga'utukusi'i.
C2	punctuation mark?	NO		"	"
C3	is it an auxiliary item?	NO		"	"
B7	inclusion mark-0?	NO	2	"	"
B8	take shorter item		hana	hana	yahaga'utukusi'i.
A2	combination allowed?	YES		"	"
A4	predictive testing	FOUND	ya	hana ya	haga'utukusi'i.
A5	end of sentence?	NO			"
A1	find the longest	FOUND	ha	hana ya ha	ga'utukusi'i.
A2	combination allowed?	YES		"	"
A4	predictive testing	FOUND	ga	hana ya ha ga	'utukusi'i.
A5	end of sentence?	NO		"	"
A1	find the longest	FOUND	'utukusi	hana ya ha ga 'utukusi	'i.
A2	combination allowed?	YES		"	"
A4	predictive testing	FOUND	'i	hana ya ha ga 'utukusi'i	.
		FOUND	.	hana ya ha ga 'utukusi'i.	Δ
A5	end of sentence?	YES			
	NEXT ROUTINE				

Figure 8.3
Example 2 (automatic segmentation).

sentence analysis from the end of a sentence. This is based on the expectation that words having a final position in a sentence are extremely limited, being confined as a rule to the function type classes *FT final verbs*, *FT final adjectives* and *FT final copulas* which offer more information about the structure of a sentence than do those occurring initially. Moreover, it seems that particles which show case, prepositional or conjunctional relationships always follow words, phrases or clauses to which they are attached, and that attributive words, phrases and clauses always stand before *DT substantives* which they modify. A complete description of the function types and essences recognized in the planned experimental system will not be given in this paper, though they are mentioned briefly in the following example for the purpose of illustrating the proposed method of syntactic analysis.

In the course of going through a sentence, predictions are constantly being generated and tested for fulfillment. As an example, let us take "nezumiganekowo (koro) sita (hanasi) ha (watakusi) wo ('odoro) kaseta". ("The story that a rat killed a cat surprised me.") Let us suppose that it has been correctly segmented through stage two of the program or automatic translation and separated into component words with their function types identified.

11.	"nezumi"	substantive FT	("rat")
10.	"ga"	*subjective particle FT*	
9.	"neko"	substantive FT	("cat")
8.	"wo"	objective *particle FT*	
7.	"(koro)sita"	*final verb FT*, takes "wo" as *objective marker*	("killed")
6.	"(hanasi)"	*substantive FT*, can take a clause of apposition	("story")
5.	"ha"	thematic *particle FT*	
4.	"(watakusi)"	substantive FT	("I")
3.	"wo"	objective *particle FT*	
2.	"('odoro)kaseta"	*final verb FT*, takes "wo" as *object marker*	("surprised")
1.	"."	*period FT*	

The procedure of syntactical analysis, somewhat simplified, is the following:

a. First of all, the prediction of 1, *end of sentence*, is stored in the prediction pool.

b. The first item to be brought into the register is a period, which fulfills prediction 1, and wipes it. It will generate prediction 3, *final particle essence*, and 2, *predicate head*, the former being placed at the top of the prediction pool, the latter at the bottom. These two predictions have what is called an "exclusion wipe mark", which causes all the predictions made by the same word to be wiped if any one which has the mark has been fulfilled.

c. The second word ("('odoro)kaseta") fulfills prediction 2, wipes both predictions in the pool, and in turn predicts 5, *object marker-A* (to be fulfilled by "wo") and 4, *subject marker*.

d. The third word ("wo") fulfills prediction 5, wipes it, and in turn predicts 6, *object master*. The content of the prediction pool is now

6. *object master*

4. *subject marker*

e. The forth word ("(watakusi)") fulfills prediction 6, wipes it, and predicts 10, *attributive substantive essence*, 9, *attributive phrase marker*, 8, *attributive adjective essence*, and 7, *relative predicate head*. *Relative predicate head* is an essence to be accepted by the so-called final-attributive forms of verbs, adjectives and copulas which modify succeeding nouns. Now the content of the pool is

10. *attributive substantive*

9. *attributive phrase marker*

8. *attributive adjective essence*

7. *relative predicate head*

f. The fifth word ("ha") fulfills prediction 4, which has what is called an "endwipe mark" which causes all the preceding predictions to be removed. The fifth word itself generates a prediction of 11, *subject master*. The content of the prediction pool is now only

11. *subject master*

g. The sixth word ("(hanasi)") fulfills prediction 11, and in turn predicts 15, *attributive substantive essence*, 14, *attributive phrase marker*, 13, *attributive adjective essence*, *and* 12, *relative predicate head*.

h. The seventh word ("(koro)sita") fulfills prediction 12, and predictions 15, 14, 13 and 12 will be wiped due to the exclusion wipe mark accompanying the *relative predicate head* prediction. Predictions newly made are 17, object *marker-A* (to be fulfilled by "wo"), and 16, *relative subject marker*. They do not have the exclusion wipe mark since the sixth word ("ha(hanasi)" can take an attributive clause of apposition, so that both predictions may be fulfilled. (If "(watakusi)" instead of "(hanasi)" occurs, for in-

stance, the two predictions will have the exclusion wipe mark since "(watakusi)" belongs to a class of substantive function types which cannot take any attributive clause of apposition, method, reason, time or place, and the relative to be inserted must be a relative pronoun of either subject or object.) The content of the prediction pool is now

17. *object marker-A*

16. *relative subject marker.*

i. The eighth word ("wo") fulfills prediction 17, wipes it, and generates a new prediction, 18, *object master*. The content of the pool is now

18. *object master*

16. *relative subject marker*

j. The ninth word ("neko") fulfills preduction 18, wipes it, and generates new predictions 22, *attributive substantive essence*, 21, *attributive phrase marker*, 20, *attributive adjective essence*, and 19, *relative predicate head*. Below these predictions is prediction 16, *relative subject marker*.

k. The tenth word ("ga") fulfills prediction 16, and reaching the "end wipe mark", wipes all the preceding predictions in the pool together with it. It generates a new prediction 23, *subject master*.

l. The eleventh and final word, "nezumi", fulfills prediction 23 and concludes the syntactical analysis of the sentence.

In the process of analysis, each word in a sentence will be assigned a) an essence which has been fulfilled by it, b) a linkage number which shows by which word it has been predicted, and c) a group number which shows to which clause in the sentence it belongs (see table 8.2).

One syntactical peculiarity of Japanese is that, unlike English, the subject of a sentence is very often omitted. Some provision must be made for insertion of subjects in English sentences when necessary. In the proposed system of predictive analysis, fulfillment of the *subject marker* and *relative subject marker* predictions is regarded as essential. If no Japanese words are found which fulfill these predictions, they are transferred together with the subject master essence, to a fulfilled essence pool with a mark to show they have no input counterpart, before they are wiped by an "endwipe mark" or after all words constituting a sentence have been tested.

Clause Level Designation of Essences Fulfilled

Word Number	Word	Essence Fulfilled	Linkage Number	Group Number
11	nezumi	subject master	10	2
10	ga	relative subject marker	7	2
9	neko	object master	8	2
8	wo	object marker	7	2
7	(koro)sita	relative predicate head	6	2
6	(hanasi)	subject master	6	1
5	ha	subject marker	2	1
4	(watakusi)	object master	3	1
3	wo	object marker	2	1
2	('odoro)kaseta	predicate head	1	1
1	.	end of sentence	INIT	0

Objects of verbs are often omitted in Japanese, but it is not necessary to supply them since objects of verbs are frequently omitted in English.

7 Transformation With Output Editing

Stage four of the program of automatic translation deals with the synthesis of the target language, in which word-order transformation is a serious problem. In brief, words which have the same group number are gathered together and within each group, transformation of word order is performed.

Subject marker, *object marker* and *relative subject marker* are omitted. *Subject master* or *relative subject master* comes first within each group, followed by *predicate head* or *relative predicate head*, and then by *object master*. For the *subject master* which has no Japanese counterpart, an imaginary substantive "X" is introduced, which has English correspondent "X".

Groups 1 and 2 of the above example will be arranged as follows:

Group 1

6. (hanasi)	*subject master*
2. ('odoro) kaseta	*predicate head*
4. (watakusi)	*object master*

Group 2

11. nezumi	*relative subject master*
7. (koro) sita	*relative predicate head*
9. neko	*relative object master*

Group 2 will be inserted immediately following "(hansi)" (Group 1), with a conjunction of apposition

"that". The inflected English correspondents will be approximately as follows:

6	story
—	that
11	rat
7	killed
9	cat
2	surprised
4	me
1	.

Conclusion

The results of preliminary manual testing of automatic segmentation on the basis of a "find the longest matching dictionary item" operation followed by "predictive testing" has given reason to believe that this program will provide a practical basis for the analysis of running kana-kanji text. Thirty-nine distribution types for Japanese have thus far been recognized, but no exhaustive classification of dictionary and auxiliary items into these types has been attempted. In particular need of further study are the problems of homographs and missing words.

Function types and essences are now under study, and experimental sentence analysis has indicated that predictive analysis should provide an effective method of obtaining the more probable analysis for a given input sentence on a single right-to-left pass. The right-to-left pass proposed, entailing as it does analysis proceeding in a direction converse to transcription, raises an important question about the syntactical nature of Japanese, and about Miller's and Yngve's hypothesis on the mechanism of temporary memory in humans.[7,8] This is a problem worthy of serious consideration.

Since the system proposed in this paper has neither been developed in complete detail nor been tested on a machine, it will be subject to various improvements as the system is further refined.

Note

This study has been supported by the National Science Foundation.

References

1. "Toyo-Kanjis, Chinese characters for daily use in Japanese," promulgated by Japanese Government in 1946 on the basis of the decision and report of Japanese Language Council; "Table of Pronunciations of Toyo-Kanjis," Japanese Government, 1948.

2. Rhodes, I., "A New Approach to the Mathematical Translation of Russian," National Bureau of Standards, Washington, D.C., Unpublished report, 1959.

3. Sherry, M. E., "Syntactic Analysis in Automatic Translation," Doctoral Thesis, Harvard University, 1960.

4. Bossert, W., V. Giuliano, and S. Grant. "Automatic Syntactic Analysis of English," "*Mathematical Linguistics ad Automatic Translation*," Report No. NSF-4, Harvard Computation Laboratory, Section VII, 1960.

5. Oettinger, A. G., "Current Research in Automatic Translation at Harvard University," National Symposium of Machine Translation, Los Angeles 1960, (to appear in *Proceedings* of the Symposium, Prentice-Hall, Englewood-Cliffs, New Jersey).

6. Oettinger, A. G., "Automatic Syntactic Analysis and the Pushdown Store," Symposium on the Structure of Language and Its Mathematical Aspects, 567th Meeting of American Mathematical Society, New York 1960 (to appear on *Proceedings* of the Symposium, American Mathematical Society, Providence, Rhode Island).

7. Miller, G. A., "Human Memory and the Storage of Information," *I.R.E. Transactions on Information Theory*, 1956, IT-5, pp. 129–137.

8. Yngve, V. H. "A Model and an Hypothesis for Language Structure," *Proc. Amer. Phil. Soc.*, 1960, 104, No. 5, pp. 444–466.

9
On the Mechanization of Syntactic Analysis

Sydney M. Lamb

This paper is concerned with possibilities of using the digital computer as an aid in syntactic analysis. Since there is some variety of opinion regarding syntax and its position in linguistic structure, I should perhaps start by giving my opinion, so that you will know what I am talking about when I refer to syntactic analysis.

There are three (and only three) types of hierarchal relationships existing among the structural units of language. They are: (1) that of a *class* to its *members* (e.g., vowel: /a/, noun: Boy); (2) that of a *combination* to its *components* (e.g. /boy/:/b/, ⟨men and women⟩: ⟨women⟩); and (3) that of an *eme* and its *allos* (e.g., /t/:[t′]). These relationships may be called hierarchical because in each of them there is one unit which is in some way on a higher level than the others.

There is a fourth type of hierarchical relationship, but it is not present within the structure of a language. It is that of a *type* to its *tokens*, and it exists as a relationship of the language to utterances or texts. Any unit of a linguistic structure is a type with relation to tokens, i.e., occurrences, of it in texts.

A listing of the kinds of hierarchical relationship to be found in linguistic structures does not, of course, constitute a complete catalogue of all relationships to be found among linguistic units since there is a type of "sibling" relationship for each type of hierarchical relationship (e.g., among members of the same class or allos of the same eme).

The eme:allo relationship is often confused with another type which in reality occurs only in diachronic linguistics. This is the relationship of a linguistic item to that which results from the application of a process to it. All of the situations in which this process relationship is used in synchronic linguistics can be better dealt with by means of emes and their allos. At the same time, there are many linguists who do not consider the eme:allo relationship to be different from the class:member relationship. That is, they erroneously speak of an eme as being a class of allos.

But the relationship of an eme to its allos is really one of *representation*. That is, the eme is *represented* by its allos on a different level. Thus the recognition of this type of relationship involves the recognition of separate levels. These levels, however, must be clearly distinguished from other kinds of levels which are set up for dealing with other kinds of relationships. Accordingly, we may use a distinctive designation, such as *stratum*.[1] In any instance of the eme:allo relationship, then, *the eme has its existence on one stratum, its allos on the adjacent lower stratum*. Every unit of a linguistic structure exists on one and only one stratum, and classes and combinations of items always have their existence on the same stratum as those items. Thus levels of the other types which are sometimes confused with strata also have their existence within a single stratum.

For most spoken languages, there are at least four structural strata. We may call these the *phonemic*, the *morphophonemic*, the *morphemic* and the *sememic*. In addition, there is another stratum, the *phonetic*, which lies adjacent to the phonemic stratum but is outside the linguistic structure. The phonetic stratum belongs to the "real world" and consists of sounds, while everything in the linguistic structure is abstract in nature and neither contains nor consists of sounds.

An indication of the kinds of features which are accounted for on the various strata is provided by the following examples:

Phonetic:	set: se.d (set, said)
Phonemic:	set: sed
Phonemic:	berk: berge (German "mountain")
Morphophonemic:	berg: berge
Morphophonemic:	gow: went
Morphemic:	go: go ed
Morphemic:	John call ed: John do ed not call
Sememic:	John ed call: John ed not call
Morphemic:	easy ly: with difficult y
Sememic:	easy ly: difficult ly

For written languages, the *graphetic*, *graphemic*, and *morphographemic* strata correspond, respectively, to the phonetic, phonemic, and morphophonemic.

The area of sememics is still being systematized, and it is not unlikely that when more sememic analysis of languages is done, it will become apparent that, for some languages at least, a *morpho-sememic* stratum, intermediate between the morphemic and sememic, should be recognized.

Any language has as part of its structure patterns according to which items are arranged on each of the strata. The term *tactics* is widely used for the analysis and description of arrangements, and the term *syntax* is traditionally used with reference to arrangements on the morphemic stratum. It is in connection with that stratum that the study of tactics has been of greatest interest in linguistics.

The items with which syntax is concerned can be of varying kinds, depending upon the school of thought. Some linguists regard the word as the basic unit of syntax; others make no syntax-morphology distinction, and we could apply the term *syntax* here also, with the morpheme as the basic unit. It is also possible to use items which tend to be smaller than words but larger than morphemes, and one unit of this kind is in fact what I prefer. I call it the *lexeme*.[2] But for purposes of this paper, let us think of syntax as being quite general with regard to the choice of the basic unit. The technique of analysis to be discussed applies for any of these kinds of items. After all, if one goes to the trouble of writing a computer program for syntactic analysis, one ought to make it as widely applicable as possible to the needs of different linguists. Indeed, the system as described in this paper, and the accompanying computer program, could also apply to the study of arrangements of phonemes or letters or syllables or morphographemes and perhaps also various non-language phenomena which tend to occur in patterned linear arrangements. In other words, it is really a system for tactic analysts in general.

At any rate, whatever unit is taken as the basis of the tactic description (word, lexeme, morpheme, or what-not) will be referred to as an *item* for purposes of this exposition.

The syntax may be completely described by a list of distribution classes of items, with the membership of each, and a list of constructions. A *construction* is characterized by specification of (1) the distribution classes which enter into it and their relative order, (2) the distribution-class membership of the constitutes. Lists of distribution classes of composite forms need not be given in the description (even though they exist), since they are defined by the constructions.

A simple notation for constructions is the following:

A B / C

"Members of class A occur with following members of class B, the constitutes belong to class C."

Illustrations of various situations and devices are given below:

(A) (B) C / C

(Endocentric construction; A and B are optional. The constitute class C is of the same brand *as the constituent class C, but of the next higher* degree. *This property may be made explicit by the technique of the next illustration.)*

A′ B / C

(A′: Members of A which are unit items, if any, plus constitutes of constructions listed above, but not those which are constitutes of this construction or constructions listed below. Constructions to be listed in order of increasing degree.)

′A* B / C

(Only certain members of A participate, as specified. No overt subclass of A set up because the restriction applies only to this construction.)

A B / C⁻

(Constitutes have more limited distribution than other members of C, as specified.)

A B* / C

(Special statement needed on relative order of constituents; e.g. discontinuous as specified.)

A (B:) / A

(The occurrence of a member of B may be repeated zero or more times).

To say that a syntactic description consists of lists of distribution classes and constructions, however, is to specify only its form. There are any number of possible descriptions for a given language which could take this form, but only a few of them are good and only one is the best. It must further be specified, then, what constitutes the best solution. Alternatively, one could specify a procedure which, if followed, leads to the best solution. This latter approach has been popular in linguistic methodology, but it tends to be unnecessarily complicated. In syntax (or tactics

in general) we can provide for good analysis and description very easily, by means of a simple definition. Taking for granted that the fundamental requirements of completeness, accuracy, and consistency are met, the best description of the syntax of a language is (naturally) the simplest. Simplicity in this area can be very clearly defined. *The simplest syntactic description is that which makes use of the smallest number of constructions*. It must also be specified that if two solutions have the same number of constructions, the one with simpler constructions is to be preferred. Thus we must now define *simple* with regard to constructions. A construction without discontinuous constituents is simpler than one with such constituents. And among constructions with different numbers of constituents, the simplest is that with the fewest constituents.

Having this sample definition of the best of all possible syntactic descriptions of a language, the analyst can use it either to show that a proposed solution is better than some alternative or, ideally, that it is better than all possible alternatives.

All valid criteria for determining immediate constituents can be deduced from the basic definitions. And most of the criteria which have been put forth by various linguists in recent years are valid in this sense. On the other hand, two principles are worthy of note as having been mentioned at one time or another without being valid. One of these is that constructions should always be binary or that they should always be binary except in the case of co-ordinate constructions having more than two members. The other, applicable only if items smaller than words, such as lexemes, are taken as the basis of the description, is that words must always be constituents.

Any procedure which arrives at a description satisfying the basic requirements is a valid one. If, therefore, one were expounding on syntactic analysis for the sake of human beings, any remarks added to the above having to do with procedures would serve only pedagogical purposes. On the other hand, if one wants to have a computer do syntactic analysis, it is necessary to specify a procedure in complete detail, since present-day machines are altogether lacking in intuition and ingenuity.

Let us now go into some general considerations relating to the application of computers to syntax, after which I will describe part of a specific procedure which I am currently working on.

The machine should use texts as its primary source of information, but it could also be enabled to ask for further information from the informant, just as human linguists do, in order to compensate for the absence of an infinite text. However, the machine will not be quite as dependent upon the informant as humans are, because, taking advantage of its capacity to process data at very high speeds, it will be able to work with much larger amounts of text than would be feasible for the human analyst. By the same token it should be able to do a more detailed analysis than is generally possible.

It need not be required in the initial attempts that the machine program be able to do the entire job of syntactic analysis. Provision can be made for it to admit failure on difficult problems, printing out the relevant data and leaving the solution up to human intelligence. Also, one can keep the initial stages simple by operating only in terms of binary constructions with continuous, obligatory constituents. Consideration of the more complicated types of constructions can be taken up at a later stage of the process.

The program should be designed to do its preliminary analysis on a fairly small portion of text (say around 5,000 items) at first, after which a larger amount can be considered for purposes of more detailed analysis. When the larger portion is brought in, its items can first be classified to the extent possible on the basis of the preliminary analysis, and tentative groupings based on the provisional constructions can be made. The data of the larger portion of text will thus be greatly simplified for the sake of the further analysis, even though some of the provisional conclusions may have to be rescinded.

For the remainder of this brief paper, let us consider just the preliminary analysis that is to be done on the first 5000-item portion of test.

In the course of the analysis, groupings of two kinds will be made. These may be referred to as *horizontal* and *vertical* groupings, or H-groups and V-groups for short. A *vertical grouping* or V-group is a grouping of items (and/or sequences of items) into a distribution class or an approximation to a distribution class. An H-group or *horizontal grouping* is a grouping of constituents of a construction (or tentative construction) into a constitute. Thus a combined horizontal and vertical grouping yields an actual or provisional constitute class. After an H-group or V-group has been made, it can be treated as a unit for the further conduct of the analysis. The term *unit* will be used from here on to refer to any item, V-group, or H-group.

But how is the machine going to make these V-groups and H-groups? Zellig Harris, in his procedure-oriented *Methods in Structural Linguistics*,[3] set up

distribution classes of morphemes before considering horizontal groupings. To do so in a meaningful way requires that items grouped together be found in identical environments extending several items on either side. It would be futile to attempt such an approach even with a machine because a corpus of truly colossal proportions would be required, and even the computer has limits with regard to the volume of data that can be processed at high speed. One must design the procedure, then, so that the sharing of certain *significant* distributional properties, rather than certain total environments, will be the criterion for combining units into the same V-group and such an approach requires that a certain amount of horizontal grouping be done first, since it is only in terms of H-groups that we can define significant distributional properties in advance of the completion of the analysis. Now it happens that there is a means of setting up H-groups which are at least usable approximations to constitutes of actual constructions, without the aid of any prior vertical grouping. This method makes use of a concept which I call the *token neighbor ratio*, or T/N ratio for short.

Any specific occurrence of an item may be called a *token* of it. The number of tokens of an item in a text is thus equal to the number of times that item occurs. Any item which occurs adjacent to another item is a *neighbor* of the latter. If two items A and B occur contiguous to each other, A at the left, then A may be called a *left neighbor* (LN) of B, and B may be called a *right neighbor* (RN) of A. The number of tokens of a given item in a text divided by the number of different right neighbors (i.e. RN types) may be called the *token/right-neighbor* (T/RN) *ratio* for that item in that text. Similarly, the ratio of the number of tokens to the number of different left neighbors (i.e. LN types) is the *token/left-neighbor* (T/LN) *ratio* for that item in that test. T/RN and T/LN are the two kinds of token/neighbor (T/N) ratios.

The first step in the analysis is to compute the two T/N ratios for every different item in the text. For a 5000-item text, this takes about eight to ten minutes on an IBM 704, depending on the number of item types present. In the course of calculating these ratios for each item, lists of right and left neighbors will be formed but they will not be saved since the aggregate of such lists would soon become very bulky and those individual neighbor lists that will be needed later can be constructed again very rapidly when needed.

The highest T/N ratios identify the points of maximum restriction on freedom of combination, insofar as such identification can be made without prior information about the structure of the language.

Table 9.1
Highest T/N Ratios of Items (Quasi-Lexemes) in a 5000-Item English Text, Excluding Ratios of Punctuation Lexemes

Item	Token Count	LN Count	RN Count	T/N
are	82		6	13.67
-s (verbal 3rd sg.)	53	4		13.25
-'s	85	9		9.44
new	8	1		8.00
have	58		8	7.25
own	6	1		6.00
Adolf	5		1	5.00
the	327	66		4.95
but	9	2		4.50
they	40		9	4.44
seem	4		1	4.00
call	4		1	4.00
Rhineland	4	1		4.00
he	75		21	3.55
be	28	8		3.50
Reichswehr	7	2		3.50
-pl (nominal pl.)	222		67	3.33
German	34	11		3.09
force	9		3	3.00
it	24		8	3.00

The process continues with consideration of the item having the largest T/N ratio. This item we may call the *current most restricted unit*, or CMRU. Later the next largest ratio will be considered, and so forth, but various ratios will also be undergoing modification to give effect to horizontal and vertical groupings treated as units, so the second highest may not turn out to be the highest after the first has been dealt with.

Table 9.1 shows an ordered list of the items (''quasi-lexemes'' in this case) having the highest T/N ratios in a particular English text, a selection from the writings of Sir Winston Churchill.[4] The neighbour class with respect to which the CMRU has the highest ratio may be called the SNC, for *small neighbor class*. It is necessarily small, moreover, its smallness has significance since the item of which its members are neighbors occurs with relatively high frequency in the text. That is, the highest T/N ratio can be the highest only by virtue of the fact that the size of T (number of tokens) is relatively large while the size of N (number of neighbors) is relatively small. It does not necessarily follow, however, that this item (the CMRU) and

these neighbors are partners of each other in a construction nor that this small neighbor class constitutes a distribution class.

In designing the procedure, one is faced with alternatives at this point. One could consider the SNC to be a first approximation to a distribution class. In this case, if it has more than one member, it would be necessary to look for the presence of certain relationships of its members to each other. Specifically, it would be necessary to find out whether any of its members can have any members of this same class as neighbors. For all those members which can, separate position classes (left to right) would have to be set up, and it is even possible (though not likely for the first neighbor class studied because of its small size) that more than one set of such classes would be present.

A simpler alternative is to let the machine refrain from making any vertical groupings at this point, waiting until more information is available as a result of the formation of additional H-groups. In general, we will want to combine units into a vertical grouping *only when they are found to share the same partner in H-H-groups which, in turn, also share the same partner in horizontal groupings of the next higher degree.* For example, if A, B, C, ... are items, and if AB and AC are H-groups, then that fact alone is not sufficient grounds for grouping B and C together (cf. *John left* and *John Smith*). But if AB-D and AC-D (or D-AB and D-AC) also become H-groups then B and C will be combined in a V-group. Even the grouping under these circumstances could be incorrect, however, so re-examination of V-groups will be necessary after further analysis has been done.

As soon as the CMRU is obtained, then, it will be combined with each member of the SNC into one or more H-groups. But since such groupings will often be incorrect, there must be provision for re-appraising H-groups at suitable later points, revising as necessary. Let us take an example. As we might expect, frequently occurring prepositions in English have relatively high T/RN ratios. Suppose that the preposition *in* in a text occurs several times, having as different right neighbors *sand*, *water*, and *the*. The H-groups *in sand*, *in water*, and *in the* will be formed. Obviously it is necessary that the last of these be rescinded sooner or later. And it will be, as soon as certain V-groups are made. The article *the* has been combined with the preceding *in* simply because the machine does not yet know that the nouns following it belong together in a V-group. (Let us leave adjectives out of the picture, to keep our example simple.) But as the process continues, these nouns will gradually be grouped together and the resulting V-groups will be treated as units. Then, if re-appraisal of affected H-groups is conducted as each new vertical grouping is made, it will eventually turn out that the T/RN ratio of *the* is higher than the ratio which led to the combining of *the* with in, and that incorrect H-group will at that point be dissolved.

It will be noted that although the procedure begins by considering immediate environments only, wider environments automatically come into consideration as horizontal groups are made.

A detailed summary of the first stage of the process follows:

Definitions

Item: ultimate constituent.

Unit: item, H-group, or V-group.

H-Group: horizontal grouping; i.e., constitute of a construction or of an approximation to a construction.

V-group: vertical grouping; i.e., provisional distribution class.

CH-group: complex H-group; i.e., H-group in which at least one partner is itself an H-group.

CMRU: current most restricted unit; i.e., the unit currently having the highest T/N ratio.

NC: neighbor class; i.e., the set of units which are neighbors (right or left) of a given unit in a given text.

SNC: small neighbor class; i.e., the NC with respect to which the CMRU has the highest T/N ratio.

Main Routine

I. Perform *A* on every different item in the text.

II. Get the CMRU and for each member of the SNC as partner form a new H-group. For each new H-group, (1) record its membership in reference list; (2) replace it in the text (each occurrence) by a unit symbol for the group (reference list permits restoration in case of later revision); (3) if it is a CH-group, go to *B*, specifying which partner is complex (if both are complex go to *B* twice). Perform *A* for each new H-group and for all units affected by the new groupings (replacing previous information now obsolete), namely (1) units occurring as neighbours of the new H-groups and (2) those members of the SAC which still have occurrence apart from the new groups.

III. *Switch*, having the values *plus* and *minus*. (Starts as *minus*, can be set *plus* by B and is reset *minus* by *IV*.) If *minus* return to *II*; if *plus* go to *IV*.

IV. Reset switch *III* to *minus*. Form new V-group(s) as indicated by *B*. For each, (1) record its membership in reference list; (2) in test, replace tokens of members by symbol for the group. Perform *A* for each new V-group and for all other units affected by the new grouping. Re-appraise all affected H-groups revising as needed; upon revision, re-appraise any affected V-groups, revising as needed. Return to II.

Subroutines

A. Determine the T/N ratios of the specified unit.

B. Split the specified complex partner into its constituents and add the CH-groups (in this form) to the list of CH-groups; let the other partner be called Other Partner. If Other Partner and either constituent of the complex partner match the two members of corresponding position of any other CH-group in the list, set the switch (III) *plus*; the third (non-matching) constituents are to be combined as a V-group.

At the time of writing, the process is operational on the computer only up to the point at which proper justification is found for making the first vertical grouping. In performing the analysis on some newspaper text from the Associated Press which had kindly been furnished by the MT group at the Massachusetts Institute of Technology, the machine reached that point after forming 31 H-groups, three of which were complex. In this text, capitalization of the following letter was everywhere segmented as a separate item by the M.I.T. group, so much of the horizontal grouping involved combining proper names (such as Poland, Gomulka, Egypt) with their preceding capitalization. The first vertical grouping consisted of *united* and *mrs*. Both had been combined with preceding capitalization, and each of the two resulting H-groups was found to have capitalization as its only right neighbor.

This is, of course, only a beginning. But it is the beginning of a system which may eventually be able to reduce the time required for analyzing the structure of a language from several years down to a few months or even weeks.

Notes

1. I have previously used the term *level*, e.g., in my paper on MT Research at the University of California, to appear in the *Proceedings of the National Symposium on Machine Translation*, but this term leads to confusion because of its wide variety of uses among different linguists. That paper explains how the stratificational system is used in MT research.

2. Even though it is defined somewhat differently from the lexeme of Bernard Bloch and Charles F. Hockett; cf. Hockett's *A Course in Modern Linguistics* (New York, 1968), Chapter 19.

3. Chicago, 1961.

4. This text consists of the first 5000 "quasi-lexemes" in the first chapter of the *Life* Magazine edition of *The Second World War* (New York, 1959). Quasi-lexemes, for this text, are the items arrived at by segmenting (1) at spaces, (2) punctuation lexemes (including capitalization at the beginnings of sentences only), (3) certain nominal (-pl, -'s) and verbal (-s, -ed, -en, -ing) suffixes, and (4) -n't and -'ll; where such segmented forms are written so that their morphemic identity in different environments is preserved, regardless of variation which might be present in a graphicemic representation.

10
Research Procedures in Machine Translation

David G. Hays

The symbolic nature of language is probably responsible for the widely held but erroneous view that linguistics is a branch of mathematics: a string of symbols "looks like" a mathematical formula; and, of course, a high-school language textbook, with its rules, looks rather like a mathematical handbook. Mathematicians are forced to adhere to the rules of mathematical systems by the high cost of mistakes (i.e., of variations from the rules). Most speakers of most natural languages never learn the textbook rules, and those who do learn them discover soon enough that the cost of breaking many of the stated rules is negligible. In fact, whereas mathematical systems are defined by their axioms, their explicit and standard rules, natural languages are defined by the habits of their speakers, and the so-called rules are at best reports of those habits and at worst pedantry.[1] There is good reason for moderate pedantry in language teaching, as G. B. Shaw—lately with the collaboration of Lerner and Loewe—preached. But processing natural language on a computer calls for precise, accurate, voluminous knowledge of the linguistic behavior of the speakers or authors whose utterances or writings are to be processed. Here we shall consider acquisition of that knowledge.

Types and Sources of Information

Any language-data processing system has a purpose. A system for machine translation (MT) is expected to accept text in some natural language, perhaps Russian, and produce text in another language, perhaps English. The output text should convey the same information as the input text; if it describes a chemical experiment, a chemist should be able to read the translation and reproduce the experiment with no more difficulty than if he had read the original report. Moreover, he should be able to read the translation as easily as if it had been written by a person fluent in the output language—for example, Russian documents should be translated into versions that might have been written by Americans.[2] Other systems—for indexing, abstracting, automatic programming, sociological or historical research, legal documentation, and so forth—have other purposes, but here we shall concentrate on a detailed treatment of machine translation.

Knowing what a system must accomplish tells the designer—clearly or not—what information must be supplied it. An MT system[3] must include a list of source-language words with their target-language equivalents; when it becomes apparent that many words have alternative equivalents, and that choosing among them causes trouble for the reader (confusing him or at least slowing him down), the designer realizes that he must supply the system with information about equivalent choice—under what circumstances each equivalent is chosen.

Even if it were possible to translate every word accurately without reference to context, readers would be dissatisfied with the results. Individual words have meanings, but it is only by putting words together in sentences and paragraphs that authors can communicate useful ideas. In a source-language text, the relationships among the words in each sentence are indicated by natural devices belonging to the syntax of the input language. Translating word by word does not carry over the indicators of relationships, since natural languages share syntactic devices only to about the same extent that they share words; there are cognate words that can be recognized in French or German text by an American who knows no French or German, and there are cognate syntactic devices that make word-by-word "translations" partially understandable, but to rely on them would make reading the MT output like solving a word puzzle. Thus the designer must furnish his MT system with information about the syntactic structure of the input language and the output language and about the correspondence between them.

For sources of information the system designer will naturally turn first to published grammars and dictionaries. A grammar[4] lists categories (of words) and rules for combining categories; it purports to describe

the syntax of its language. A dictionary lists words and specifies for each the categories to which it belongs; each entry also contains a discussion of the meaning of the word or a list of its equivalents in a second language. Taking grammars and dictionaries together, it should be possible to read and write grammatically correct sentences, translating each word accurately. Unfortunately, published grammars and dictionaries of the best sort are inadequate, even though they are vast compilations based on the prior original work of many linguists.

The largest dictionaries are intended to meet the needs of laymen, not of professional linguists; consequently, they omit reference to many categories that the layman can either recognize intuitively or disregard when he sees an unfamiliar word in text. The most detailed grammars are written for linguists who, recognizing that new words can be added to existing categories, make no attempt to list every word in every category. In general, until computational linguistics was conceived, no one needed a fully detailed account of any language for any purpose. Now that the need has arisen, new data must be collected and analyzed.

There are qualifications, of course. Fully detailed accounts of language have scientific value for linguistics, since they permit more exact tests of theory than gross statements about general tendencies could support. Furthermore, the major grammatical treatises dealing with Western languages—English, Russian, and others—contain many lists of words with special properties; these lists can be used to elaborate dictionaries by noting, in the dictionary entry for each word on the grammarian's list X, that the word has property x. But even a combination of information from multiple existing sources does not lead to a final, complete dictionary and there is still information to be gained from research.

The linguist has two sources of information beyond published studies. He can consult persons who speak the language, called *informants*. He can also study *text*, either written in the language or transcribed from conversations spoken in it. Of course, the published studies go back to exactly the same sources in the end. The two kinds of data sources can be used in tandem, with the informants serving as *editors* who comment on the text. Moreover, it is possible to obtain or create parallel texts in two languages, perhaps one known and one unknown, or one the input and the other the output of a proposed translation system.

The traditional methods of linguistics are based on the use of informants, or the alternate use of text and informants.[5] For non-Western languages, at least, it is fair to say that the success thus far achieved in scientific linguistics is the result of rich technical development and careful application of the informant method. Western languages have been studied by text methods and also with informants; often the linguist serves as his own informant when he is studying his own native language. The largest, most detailed grammars now in existence are the text-based grammars of Western languages, and it seems inevitable that text must supersede the informant when the details are to be filled in, simply because no one knows every particular of his language. Certainly no one knows any modern language, well developed as a medium for scientific and scholarly communication, in all its specialized ramifications. The informant learns his language by formal training and, more importantly, by constant exposure to its use. He cannot repeat to the linguist what he has never seen or heard. A sufficiently diverse set of informants would serve for any language, but the practical difficulties are obvious.

Moreover, data collected by textual research have a certain validity that data obtained from informants can never possess. An MT system, or any other automatic language-data processing (ALDP) system, will be called on to process segments of text from a definable stream. Predictions about the nature of that stream can be made, by the ordinary logic of statistical inference, from samples of it. Predictions can also be made from the responses of informants, but then the logic of inference must take into account the informant as a device that gathers information, summarizes, forgets, distorts, and reports.[6] The linguist should wonder whether he could not design a procedure that would process the same material as the informant more accurately and with less distortion.

The question of procedures for linguistic research always founders in discussion of the informant's intuition. The informant is more effective than a computing machine as a device for linguistic data reduction, according to this argument, because he understands the text to which he is exposed. The argument seems to come down to two points. First, the informant has a rough-and-ready grammar for his own language, which he uses as a framework on which to hang whatever new grammatical details come to him in reading or listening to new material. Second, he uses semantic analysis of text in deciding what its grammatical structure must be. As we shall see, the first point does not differentiate between

computers and informants, since the linguist establishes some sort of grammatical framework at the very beginning of his research and commits it to machine memory; the framework may come from specific knowledge of the language to be studied or from a theory of linguistic universals, but it is essential. The second point is more significant: Can the grammatical structure of a language be determined without reference to its semantic structure? If this question receives a positive reply, as it does from some but not all linguists,[7] then *should* grammar and semantics be kept apart? We cannot even begin to answer this question until we have looked into the nature of grammar, in following sections. In any case, research procedures based on text can be formulated with whatever admixture of informant intuition is considered appropriate.

The invention of techniques using text alone, with no help of any kind from informants, is one of the most exciting problems in linguistics today, and stimulation of work along this line may prove to be the most important contribution of the computer to the science of language.[8] The problem is to give an adequate characterization of the object of grammatical research without reference either to the intuitions of the informant or investigator or to the infinite *corpus* (body of text) that would resolve all questions if it could be written and studied.[9] (Grammatical statements often have the form *Item X can—or cannot—be used in context Y*. Such a statement would have an obvious empirical interpretation with reference to an infinitely long text in which everything occurred that could occur.)

Edited text can be used with less inventiveness; it is therefore a more practical material for the investigator who wants immediate results in the form of at least approximate knowledge about the speech habits of authors using a certain natural language. Given a text, editor informants can be asked to translate it, to paraphrase it, to describe the grammatical relations within each of its sentences, and so on.[10] The editor certainly uses his ideas about grammar, his semantic understanding of the text, and all his "intuition," in this process. The linguist's task is to generalize and formalize the informant's intuitive analyses of single sentences into a description of the language as a whole, testing along the way for consistency, completeness, and simplicity.[11]

This discussion, therefore, is largely devoted to research methods based on text. Informant-centered methods are well described in the current literature, and text-based methods have definite advantages.

Text-based methods also have disadvantages that must not be forgotten. A large amount of text has to be processed before the investigator collects an adequate number of occurrences of any but the few commonest words or constructions. The cyclic method, to be described below, avoids this difficulty so far as possible by using a computer for much of the processing work. Another problem is the influence of the general environment on the content of any text. Caesar never wrote about television, yet no linguist would believe that the rules of Latin grammar prevented him. If there are no "octagonal whales" in our text, is it because of grammatical rules or not? The answer can only be that the distinction between grammatical rules and rules of other kinds is somewhat arbitrary, and will often be decided in terms of formal criteria without help from intuition. Only a dogmatist invariably knows a grammatical regularity when he sees one.

Grammar

Grammar is a branch of linguistics. In a coherent treatment of the science or of a language, the study of grammar follows discussion of phonetics and phonemics—dealing with the sound system by which language is communicated orally—and of graphetics and graphemics—dealing with the writing system. Grammar itself has two main branches, morphology and syntax. Beyond syntax lies semantics, which will be considered later.

Morphology has to do with the analysis of words and *forms* of words. In some but not all languages the word forms that occur in text can be subdivided into repetitive fragments; that is, relatively few fragments combine and recombine in many ways to yield a large vocabulary of forms. In an MT system it is economical to avoid storing repetitive data if they can be reconstructed by a simple program from a smaller base; hence storage of fragments instead of full forms is usually advocated by system designers.[12]

More than economy is involved, however, since morphological analysis lays the foundation for syntax. Typically, the forms of a language can be segmented into prefixes, stems, and suffixes. For example, *inoperative* = *in*+*operate*+*ive*. A single form can consist of no prefixes or one or more prefixes, one or more stems, and no suffixes or one or more suffixes.

It seems to be a universal feature of natural languages that if forms can be segmented, some of the segments are involved in syntactic rules. Thus *operate*

is a verb, but the *-ive* suffix converts it into an adjective; *boy* is a singular noun, *boy+s = boys*, a plural noun. In Latin, Russian, and other languages, noun forms can be segmented into stems and case-number endings; the case endings are involved in syntactic agreement with verbs, prepositions, etc.

The morphological classes in a language are classes of prefixes, stems, and suffixes. The classification is established by noting that some stems occur with certain prefixes and suffixes attached, but not with others. A noun stem, morphologically, is a stem that occurs with suffixes belonging to a definite set—the noun suffixes of the language. A verb stem is one that takes verb suffixes, an adjective stem one that takes adjective suffixes, and so forth. Prefixes are sometimes peculiar to nouns, verbs, adjectives, etc., and sometimes are attached to stems in categories that cut across morphological parts of speech.

A form, consisting of certain definite segments, can be assigned to a morphological *form class* according to the class memberships of its components. This classification of the forms in a language is the eventual contribution of morphology to syntax; any procedure for syntactic research can begin with form classes rather than with individual forms.

Syntax has to do with the analysis of sentences and the relations that obtain among the forms that occur in them. The structure of a sentence can be described in several ways; the theory of *dependency*, as used here, is familiar to anyone who has studied grammar in school. Tesnière elaborated the concept,[13] Lecerf contributed to the theory,[14] and the present author and his colleagues are using it in studies of Russian.[15] According to dependency theory, a partial ordering can be established over the occurrences in a sentence. One occurrence is independent; all the others depend on it, directly or indirectly. Except for the independent occurrence, every occurrence has exactly one *governor*, on which it depends directly. The diagram of relations among occurrences in a sentence is a tree, an example of which is given in figure 10.1.

The syntactic structure of a sentence also includes a typification of each dependency link. Each dependent serves some definite syntactic *function* for its governor; one governor can have several dependents, all serving distinct functions, but it can have only one dependent serving any single function. (Of course, a given function can be served by several conjoined occurrences or by two or more occurrences in apposition.)

A sentence printed on a page is a linear array of letters, marks of punctuation, and spaces; morphological analysis converts it into another linear array, this one consisting of occurrences of segments grouped into forms and punctuated. If a sentence has a syntactic structure, it must be deducible from this array. The *indicators* that are available in natural language, the grammatical devices mentioned earlier as requiring translation along with the "words" in a text, include inflection, function words, occurrence order, and punctuation (in written language) or intonation (in spoken language). The use of these indicators is controlled by syntactic rules.

Inflection is used to show that a word that can serve several alternative functions in the language is in fact serving one in particular in this occurrence. For example, in Russian, a noun is inflected to show case: nominative when it functions as subject of a verb, accusative when serving as object, etc. Inflection is also used to show concord; a Russian adjective agrees with the noun it modifies in number, gender, and case, although one would not say that it has different functions corresponding to the different noun genders.

Function words are used in many languages; they have little or no *meaning*, in the ordinary sense, but serve only as indicators of syntactic structure. Prepositions, for example, differentiate functions more precisely than the case system can do; Russian has half a dozen cases and about fifty prepositions.

If each sentence contained no more than one word capable of governing any given function word or inflectional category, occurrence order would be al-

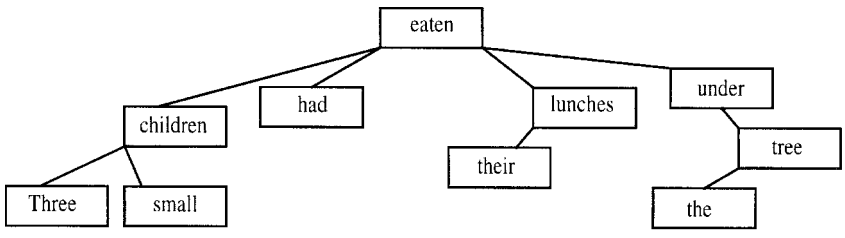


Figure 10.1
Dependency structure.

most irrelevant; an accusative noun in Russian, for example, might be recognized in any position as the direct object of the verb in the same sentence but for the fact that accusative nouns can serve other functions and other potential governors can occur along with a verb governing the accusative. (There is also the problem that some noun forms are ambiguous; they may be accusative or some other case.) Some prepositions, for example, govern the accusative, and a noun, a preposition, a verb, a comparative adjective or adverb, etc., can govern a genitive noun. Occurrence order therefore has to indicate which of several potential governors is actually served by a given occurrence. Occurrence order even differentiates functions; in English, the subject of a verb ordinarily comes ahead of it whereas the object ordinarily follows, and the two are not morphologically distinguished except when one agrees with the verb in number and the other does not.

Punctuation serves sometimes to enforce a connection (as in hyphenated combinations), sometimes as a barrier to connection, sometimes to set off a semiparenthetic portion of the sentence. Intonation, historically the ancestor of punctuation, serves somewhat the same indicative role in spoken language.

One further kind of indication is given by word-class membership. The inflected forms of a word often share properties that help to indicate sentence structure. For example, words that govern objects (a syntactic function) can be taken as a class, and words that govern, as objects, accusative nouns are a subclass.

The *syntactic type* of a complete form is given by listing the functions that it can serve for all possible governors, the functions that possible dependents can serve for it, and the properties involved in agreement with potential governors or dependents. This information takes into account word-class membership and inflectional category; each function word in a language is likely to have a syntactic type peculiar to itself. Represented in a glossary by a grammar-code symbol, the syntactic type of a form is its whole contribution to the indication of the structure of any sentence in which it occurs.[16]

We can now see what the grammatical part of a machine-translation system must do: Using the indicators of a natural language—syntactic types, occurrence order, and punctuation, in conjunction with syntactic rules—the system must determine the structure of each input sentence, i.e., the dependency links and their functional types. Then, given the structure of a sentence, the system must find devices in the output language with which to indicate that structure. On the input side, there may be ambiguities; sentence-structure determination can end with more than one possible interpretation of a given sentence. Semantic analysis, as we shall see, can reduce this ambiguity in many or most cases. On the output side the system should be designed to avoid introducing new ambiguities, although it seems likely that goal can never be fully accomplished.[17]

Semantics

Sounds or letter sequences indicate what forms occur in a text. Grammatical devices indicate what syntactic relationships obtain among the form occurrences. And the words and syntactic relationships in a text indicate its meaning. The concept of syntactic structure can be formalized, perhaps as outlined in the preceding section, and the grammatical devices of a language inventoried. When we turn from syntactic theory to semantic, we face a blank wall; no adequate formulation of semantic structure is available today. Nevertheless, we are already able to survey at least some of the problems with which a semantic theory must cope and to offer at least some specific characteristics that a semantic theory must possess.

The segmentation of forms into prefixes, stems, and suffixes does not imply that those segments are the units to be translated. As we have already seen, some segments are used in the input text to indicate syntactic relationships, and it is those relationships that have to be translated, by means of appropriate indicators in the output language, not the segments themselves. Other individual segments do in fact have to be translated, but it is sometimes most convenient to translate combinations of segments within one form and occasionally combinations that include segments of several forms. The choice of units is connected with the determination of meanings.

Much evidence goes to show that the words of natural languages are ambiguous—i.e., have multiple meanings.

In translation, as from Russian into English, French, German, etc., a given Russian word may have many different equivalents in each output language, and its English equivalents may not translate unambiguously into French even if the correlation with a Russian word is known.[18]

Monolingual dictionaries give multiple definitions for individual words, and, as Kaplan has shown, native speakers given context can "resolve the ambiguities" by assigning dictionary definitions to form

occurrences.[19] Here it is only the fact of inter-informant reliability that is convincing; no one informant could convince us that a real difference exists between two dictionary definitions of the same word, but if several informants, consulted independently, agree that occurrences A, B, C, . . . , take the first definition, whereas occurrences X, Y, Z, . . . , take the second, the difference clearly exists for speakers of the language. In conducting a test of this type it is necessary, of course, to remember that informants can be ignorant of distinctions that other users of their language make with regularity and precision. On the other hand, dictionaries are not infallible either, and they undoubtedly contain distinctions that are not known to speakers of the language, at the same time missing distinctions that are widely known.

A third line of evidence suggested by Harris[20] is that words with the same meaning should occur in the same range of contexts (have the same distribution, in the linguistic sense). It follows that a word with two meanings should occur in two distinct, separable ranges, i.e., its distribution should have distinguishable parts corresponding to the two meanings. All known suggestions for the resolution of ambiguity in ALDP systems, as well as all suggestions conceivable in computing systems limited to textual input, are based on this notion. Our point for the moment is simply that if a word occurs in two distinct ranges of contexts, and grammatical theory does not explain its distributional peculiarity, then semantic theory must be adduced.

The evidence that establishes multiple meaning as a linguistic phenomenon does not provide for determining exactly how many meanings each word has and how the boundaries are to be drawn. Informants may agree that a certain word has two meanings, yet not agree on its meaning in certain contexts, or a large group of informants may agree that it has two, while a subgroup divides one meaning into two, making three altogether. Translation into one language may require two equivalents for a certain word, into another three, and it may be argued that some of the equivalents differ only stylistically or syntactically. Distributional evidence likewise ranges from strikingly clear to suggestively vague. In point of fact, a search for precision by any of these methods is likely to be thwarted, since all of them are indirect.

The three lines of evidence so far mentioned are all linguistic, whereas semantics must deal with the relations between language and reality, or, if reality is elusive, cognitive and cultural elements. Reality, as far as we now know, is infinitely complex, and languages, like science and all of culture, are finite. On a smaller scale, it would be nonsense to claim that the English word *hat* has as many meanings as there are, have been, or will be hats (headgear) in the world. All those hats are simply different referents for a single meaning of the word. No more does *bird* have as many meanings as there are species or varieties of Aves; one meaning covers them all. If a badminton bird is something else, it is because the culture has an organization independent of the language, and egg-laying birds are culturally differentiated from feathered hemispheres at a very deep level. It is *not* primarily a linguistic fact that the properties characteristic of birds (robins, canaries, etc.) and the properties characteristic of (badminton) birds are practically nonoverlapping. This fact pertains to the culture, to the cognitive systems of persons bearing the culture. Reality influences culture, and culture influences language; better said, the nonlinguistic part of culture influences the linguistic. Hence linguistic evidence, though indirect, can be used in the study of meanings.[21]

Each meaning of a word, then, is a cultural unit corresponding to a segment of reality that the culture regards as relatively homogeneous. A formal theory of meaning will have to go further, relating meanings to one another and giving an exact theoretical account of "relative homogeneity." One possible method is to list properties that the culture employs in forming concepts of reality; then a segment is relatively homogeneous if it can be distinguished from other segments by many properties but only subdivided by a few. Or it may be necessary to recognize that some properties are more significant to a culture than others and to decide homogeneity on the basis of the significance of the properties that isolate a segment as against the significance of those that cut it into subsegments. As yet we can say no more than this about the formal analysis of ambiguity.

Another semantic problem that we must consider is the calculation of the meaning of a sentence from the meanings of its constituent words or word segments. Syntax is needed in language to reveal semantic connections among the parts of sentences. In most sentences, for example, interchanging the subject and object of a verb alters the meaning of the whole in striking fashion; when the propaganda organization of a dictatorship announces that "Nation A has committed acts of aggression against nation B," interchange of A and B in such an announcement would be treasonable. Semantic relations are not identical with syntactic relations, however, and the

same problems of identifying distinct meanings and resolving ambiguities arise with relations that we have already considered for words. We can begin with syntactic functions and attempt to determine how many different semantic relations can be indicated by each function. As before, we can use textual methods in research, but we must remember that these methods are indirect; the meaning of a syntactic function is a kind of relation that is identified by the culture and isolated from other kinds of relations.

With a theory of semantics in view, we can return to the problem of isolating translatable units in language. For some—but not necessarily all—of the segments that he isolates in a language by morphological methods, the linguist can determine one or more independent meanings. He certainly excludes those segments that serve only to indicate syntactic relations, since he must deal with them separately. He next considers word forms made up of combinations of segments, always excluding segments of purely grammatical (syntactic) significance. If the meaning of a word form can be calculated from the meanings of the component segments by a standard rule—i.e., a rule that holds for many forms in the language—then the segments are translatable units. If not, the form itself must be taken as a unit for translation.[22] Thus there are meaningful morphological relationships in language as well as meaningful syntactic relationships; each permits determination of the meaning of a combination from the meanings of the parts. Again, the linguist must examine combinations of forms in the language, testing whether the meaning of the combination can be calculated from the meanings of the forms and the syntactic functions that tie them together. When a combination appears with meaning that cannot be calculated in this fashion by a general rule, the combination must be treated as an *idiom*, or translation unit larger than a single form. The general rules correspond to semantic relations one to one; a single rule may not suffice for all occurrences of a single syntactic function and therefore would show multiple meaning: the syntactic function can indicate more than one semantic relation, each associated with a rule.

Consider now the requirements of the semantic part of a machine translation system. Taking sentences with known syntactic structures as input, the system identifies the translatable units and determines both the meaning of each unit and the semantic relations that obtain among the units. Then, given a representation of the meaning of each input sentence, the system must find words and semantic relations in the output language that express the same meaning. The output-syntax system operates on the results to produce sentences in which the meaning is indicated as clearly as possible. There may be ambiguities, of course, and two or more possible meanings may be discovered for a single sentence. Until semantic theory and research have progressed and additional systems are elaborated to go beyond semantics, the MT program can only offer alternative output sentences or a single sentence with the same ambiguity as that of the input.

[. . .]

Semantic Recognition

When the syntactic structure of a sentence has been determined, and the minimal units with independent meaning have been identified, the meaning of each occurrence in the sentence and the nature of the semantic connections among them have to be determined. Work in this field does not yet enable us to describe procedures of proven effectiveness, but some suggested methods can be reported.

It may be possible to assign meanings to semantic agreement classes.[23] In that case, a table could be used much as a table of dependency types is used in SSD. An entry would consist of a pair of semantic-class symbols and an indication of a syntactic function. The question would then be, *Can an item of this class serve that function for an item of the other class? If so, what is the semantic relation between the two?* For example, *Can the name of a person serve as the (syntactic) subject of a verb of communication?* The answer would be, *Yes; the person is actor*. The classes in this example, are, of course, not necessarily those that would appear from empirical research.

Following this plan leads to success if one and only one meaning of each occurrence in a sentence agrees with the meanings of neighboring occurrences and if each syntactic connection is resolved to a unique semantic relation. If the sentence has more than one possible syntactic structure, semantic disagreements may rule out some or all of them. The semantic classes and agreement rules therefore have to be designed to determine a unique meaning for each word occurrence and each syntactic relation in every sentence, to eliminate all but one possible structure for each sentence, and to assign to every (intuitively acceptable) sentence a semantic description that can be translated or otherwise manipulated to the satisfaction of whatever external criteria are applied. As

yet there is no evidence that any semantic agreement system can approach this design standard. [. . .]

Another proposal is to organize the vocabulary of *meanings* hierarchically; formally, the organization would be a lattice.[31] Choosing any set of meanings arbitrarily, we find that there is some set of meanings in the hierarchy that includes all of them; in fact, there may be many sets with that property, one having a smaller count of meanings than any of the others. The occurrences in a sentence have meanings that can be found in this hierarchical system, but each occurrence can have one or more meanings. In a three-occurrence sentence, for example, let the first occurrence have two meanings: 1*a* and 1*b*. Let the second have two meanings also—2*a* and 2*b*—but let the third occurrence be unambiguous—call its meaning "3." We must choose one meaning for each occurrence; we can choose: 1*a*, 2*a*, 3; 1*a*, 2*b*, 3; 1*b*, 2*a*, 3; or 1*b*, 2*b*, 3. Trying each set of meanings in turn, we learn the size of the smallest class in the hierarchy that includes all meanings in the set; e.g., how large is the smallest set that includes 1*a*, 2*a*, and 3? We thus obtain four quantities, one associated with each set of equivalent choices for the sentence, and we take the set of equivalents associated with the smallest of those quantities, since semantic *homogeneity* is to be expected in an ordinary text. Ties are possible, however, which lead to semantic ambiguities. The difficulty with this model is that syntax cannot be combined with it in any obvious way. In fact, the proper solution to semantic problems could be a combination of the two methods that we have described—the first takes advantage of local context, the second uses broad context. The ambiguities not eliminated by one might then be resolved by the other.

[. . .]

Translation of Words: Semantics

Standardization of Equivalents Let us next turn to a more general view of the problem of pairing meaningful units in the input and output languages. Suppose that we must begin with a new pair of languages, for which the only available information is that contained in published dictionaries. We are to proceed by the cyclic method, processing text in successive batches. At first the relatively few most common words in the language will dominate our lists of new forms; many forms found in the first batch of text will occur frequently both in that batch and in succeeding batches. Other words in the first batch, and most new words in later batches, will be rare. The basic plan for assigning equivalents can therefore reasonably change as the number of batches processed increases.

The first batch is prepared, and an alphabetic list of the forms that it contains is made. It is convenient to collect forms into groups, when the input language is highly inflected, since the translational equivalents of different inflected forms of the same word will usually be identical (except for what we will regard as grammatic variations). Now each form or word in the text-based list can be looked up in a published bilingual dictionary and the equivalents listed for it copied into machine storage. The list of forms, each accompanied by one or more equivalents, is an initial glossary.

The next step is to list the first batch of text, with its machine translation in parallel. But of course the translation is merely a statement—for each form occurrence—of the equivalents shown for that form in the initial glossary. Now the editor informant, who should be well acquainted with the subject matter of the text, selects an equivalent for each occurrence in the first batch. He can mark one of the listed equivalents, write in a new one, or identify an idiom. His marks are keypunched and correlated with the machine-stored text and translation.

The number of times that the editor selected each equivalent for each form in the first batch is easily determined by an automatic process. Then the equivalents of each form, including those inserted in advance and those added during editing, can be ordered automatically by frequency of choice. Although separate records must be kept for each *form*, the ordering should be done for each *word*; i.e., the equivalents of all forms of a word should be kept in the same order. Now the typical glossary entry consists of a form and an *ordered* set of equivalents, together with a code symbol if the form was used idiomatically.

Treatment of the second, third, etc., batches of text proceeds in the same fashion, with two modifications. Beginning with the second batch, the editor is instructed to use the first equivalent listed with each occurrence whenever it is substantively accurate. When the second and subsequent equivalents of a word are never used by editors working under this instruction, the linguist can be sure that the alternatives differ only stylistically in the stream of text that he is processing. On the other hand, if one (or more) of the alternatives is still used, its meaning is substantively distinct from the meaning of the most frequent equivalent, and the linguist can look for contextual indicators of the difference.

The second change in procedure is omission of the first step—the insertion of tentative equivalents found

in a published dictionary. This change is justified when the average number of occurrences of each new form is small enough; generally, the reasonable level is an average of two occurrences in a batch of text. There are several reasons for this modification. For one thing, the new words are hard to find; some are in the dictionaries, some are not. For another, the proportions of cognates and proper names increase. And for another, the equivalents obtained grow less and less reliable. Altogether then, it seems best to add equivalents only during editing, once a good glossary has been developed.

Input-language Inflection and Choice of Equivalents Most words, or form groups, have uniform translations, but not all. Some Russian verbs have one English equivalent in their nonreflexive occurrences, another (not passive of the first, which would be considered the same equivalent, modified grammatically) in reflexive occurrences. Some nouns have one equivalent in the singular, the same equivalent or another in the plural. These exceptions to the general rule must be discovered and taken into account. The procedure is simple and straightforward. A file of equivalent-choice data, tallied by form and grouped by word, is required. With each form, the file must include a grammatic description. The procedure is applied to each word that satisfies three tests: (1) at least two forms have occurred; (2) at least two nonidiomatic equivalents have been chosen; and (3) enough occurrences of the word have been processed for reliable conclusions to be drawn.

The procedure is to sort the forms of a word into grammatic categories and for each equivalent test whether it occurs equally often in each category—that is, in proportion to the total number of occurrences of the word in each category. A statistical test for nonproportionality should be applied; although the satisfaction of its underlying assumptions is by no means clear, the chi-square test is perhaps appropriate.

The exceptional words can be listed and the findings installed in the glossary so that when new text is processed only applicable equivalents will be printed with each form occurrence.

Equivalent Selection by Contextual Criteria Perhaps the majority of substantive equivalent-selection problems can be resolved by reference to grammatically related occurrences in context. A verb, for example, may take one equivalent or another depending on its subject, its object, or a modifier. An adjective is most likely to be influenced by properties of the noun it modifies, since adjectives usually occur without dependents. In any event, establishment of rules for the determination of equivalents must include analysis of context.

It appears that analysis of related occurrences should be organized by kind of relation—i.e., by grammatic function. The procedure would be applied to each word with two or more equivalents, both applicable to at least some forms, when sufficient occurrences had been processed to permit anticipation of reliable results. All occurrences of the multiple equivalent form are collected; the required file of information includes *what* words were related to the given word, and with what function, as well as the choice of equivalent that was made.

The analysis then takes one function at a time; since every occurrence has a governor, let us start with that. A particular word can serve different functions for its governor in different occurrences: A certain noun can occur in one place as subject of a verb, in another as object of a preposition, etc. If the multiple-equivalent word that we are studying has only one equivalent in each kind of relation that it enters as dependent, the analysis is complete; the problem shifts, as it were, from semantics to grammar. If there is any kind of relation in which the word has two equivalents, the analysis continues by examining each *word* that governs the multiple-equivalent word. If a certain word as governor always—in the processed text—implies a certain translation for its multiple equivalent dependent, that fact is recorded; if the same can be said of every governor, the evidence suggests that choice of equivalent depends on type of governor. If not, a summary statistic can be computed—that is, the percentage of occurrences for which the correct equivalent can be selected by inspection of the governor.

The summary statistic is computed in the following manner. Let $E_1, E_2, \ldots, E_n$ be the equivalents of a word, and $C_1, C_2, \ldots, C_n$ be criterion classes. Considering only one class of related words (e.g., governors), assign each related word to class C_i if E_i is chosen more frequently than any E_j, $j \neq i$ when the related word is present. (In case of ties, assign at random.) Then, assuming that E_i is chosen when a word in class C_i is present, the summary statistic is just the number of correct choices divided by the number of occurrences of the multiple-equivalent word. This fraction, which we can call p, is at least as large as the ratio of choices of E_i, the most frequent equivalent, to occurrences of the word under study; and p is no larger than unity, which would indicate complete accuracy.

In fact, since the number of errors with each distinct related word is limited to $(n-1)/n$ times the number of occurrences of that word, the expected value of p must increase with the number of distinct words related in the given way. The sampling distribution of p, under the hypothesis that the classification of related words is irrelevant, still has to be calculated, and from it, parameters for normalizing p could be deduced.

Continuing our analysis of a multiple-equivalent word, we would examine words in each possible grammatic relation to it, calculating p, or a normalized variant p^*, for each relation. The relation for which p^* had highest value would deserve the attention of a linguist, since a few errors that might prevent p^* from reaching unity could be due to careless editing. If no value of p^* was high enough to be useful, the automatic analysis would have to continue by combining criteria. Harper, for example, working with a less formal method of analysis, used both governors and objects of prepositions to determine the equivalents required.[33] There is no certainty, of course, that the governors and dependents of an occurrence determine its translation, but it seems plausible that they will often do so.

When the criterion classes can be discovered, their members have to be marked in the glossary. A generalized semantic-recognition program can use these marks to select meanings, and thus equivalents, for occurrences of the words to which the method is applicable. So far it has been assumed that criterial classes are defined independently with respect to each multiple-meaning word. That plan would eventually call for the storage of a vast amount of information. However, it is desirable to reduce the requirements, or at least to be assured that redundant information is not stored. Furthermore, the criterial classes can reasonably be interpreted as semantic classes only if they are relatively few in number and if word meanings fall into classes that allow use of the same criteria with all members of any given class. The question is therefore whether criterial classes formed in different ways are identical. With finite text no two classes are likely to have exactly the same members, but a degree of overlap exceeding random expectations would be evidence of relatedness. Two classes, criterial for selection of the meanings of two different words, are the same class if every word belonging to one also belongs to the other; no matter how large the corpus, there is always some chance that a sentence will occur in which a member of one class gives an incorrect result when treated as a member of the other class. This possibility must be eliminated before a sound model for statistical inference can be formulated. If "exceptions" are allowed, an alternative formulation is to coalesce two classes whenever the cost of storing and manipulating one list with known exceptions is less than the cost of storing two lists with no exceptions. To make this alternative attractive, an intuitively acceptable estimation of the relative costs must be made.

Syntactic Research

Problems of morphology come up in the study of non-Western languages and even in work with Russian or English when it becomes necessary to cover all details with a uniform scheme, but problems of syntax are much more significant in current work on well-known natural languages. In this section we shall assume the existence of a complete and unchangeable morphological description of the subject language; working on that assumption, we consider several plans for syntactic research.

After a sentence-structure-determination program has found all possible structures for a sentence, an editor informant examines them and chooses the correct one if it is listed. Errors in grammatical classification or tabulation of dependency types, as well as failures of the syntactic theory, can cause the SSD program to miss the correct structure; in that case, the editor must add the structure he desires to those listed. His notes, covering both connections and functions, are keypunched and collated with the stored output from SSD in preparation for analysis.

The sentences for which the editor wrote structures not found by the SSD program must be processed first, since they reveal major gaps in the system. The first step is to test for projectivity. The program examines each connection in a sentence and determines whether every occurrence between the members of the connected pair derives from one or the other of them. If not, it marks the connection as nonprojective; such a connection needs further study by a linguist.

In the SSD program considered above, the establishment of connections in a sentence is in a fixed order, which we can call the *recognition order*. The primary sequencing variable is the size of the subtree that results from a connection. Two subtrees are assembled only by connecting the independent element of one with the independent element of the other; dependents of a given node must be attached first on one side and then on the other. For several

reasons, it is necessary to alter grammar-code symbols as connections are made; the alterations follow instructions in the table of dependency types. Thus at the time any connection is made, the grammar-code symbols of the occurrences to be connected are the result of all prior connections in which they have participated. When the correct structure of a sentence is not found by the SSD program and if the structure is projective, it must contain a connection that is impossible according to the existing system. To ascertain the cause, the SSD operation has to be repeated.

A controlled SSD program can be used for this purpose. The control is based on knowledge of the correct connections in the sentence; these connections are taken in recognition order and tested in turn against the table of dependency types. Alterations in grammar-code symbols are made as they were originally. When the "impossible" connection is reached, the SSD program has constructed a list of all grammar-code symbols assigned to each of the two connected members as a result of the alterations keyed by prior connections. Considering all possible pairs of these symbols, the program determines whether some alteration is responsible for the failure to find a suitable entry in the table of dependency types. In other words, if the impossible connection could have been made but for the alteration of a grammar-code symbol at the time of a prior connection, the alteration can be blamed for the failure to produce a correct structure for the sentence, and the relevant information must be printed out for a linguist to examine. The linguist can decide whether to change the alteration instructions, change an entry in the dependency table so that the latter connection can be made in spite of the alteration, etc.

If no alteration is responsible for the failure, the difficulty is in the grammar-code symbols, in the dependency table, or in lack of an alteration that should have been made. What is possible at this stage depends somewhat on the organization of the table.

In the simplest case, the table is a list of pairs of full grammar-code symbols. The symbols belonging to any pair of occurrences that have to be connected can be added to the table, but a screening process must eventually be carried out to avoid recognition of an excessive number of false structures for sentences in the future text. The screening program can be exemplified in Russian. In this language there is a morphological category of nouns; noun forms are morphologically subclassified by case. When enough connections between noun governor and noun dependent have been recognized in text, the screening program can detect that the case of the dependent is relevant to the function it serves, whereas the case of the governor is not. To reach this conclusion, the program must consider all morphological categories, testing for morphological diversity within each functional type; finding that every noun dependent serving a given function for a noun governor is in a certain case, the program can conclude that the case of the dependent is relevant. On the other hand, the program finds that a noun governor in any case can take a dependent noun with a given function; hence the case of the governor is irrelevant. The screening program also builds word classes. In a statistical sense, the nouns that can serve a given function when they occur in a given case are lexically diverse—many different nouns are found as dependents with any given function. Governing nouns, by contrast, are lexically restricted; the number of different nouns that govern the instrumental case, for example, is much smaller than the number expected by chance if every noun is capable of governing instrumental nouns. The statistical evidence proves the existence of a syntactic class; membership in the class is proved only by occurrence in the defining context—in the example, a noun is added to the class when it occurs as governor of an instrumental noun serving a particular function.

Once syntactic word classes are established, the organization of the dependency table can profitably be elaborated. Each grammar-code symbol will be cut into three parts: the morphological part of speech, other morphological properties, and syntactic word-class memberships. When two occurrences are said to be connected and the dependency table cannot connect them, the parts of the grammar-code symbols can be tested in turn. Continuing the previous example, let us suppose that both occurrences are nouns. First, parts of speech are consulted. Second, given the function named by the editor, the relevant morphological properties are sought in the table. If two occurrences of the given morphological types cannot be connected, an entry must be added to the table (but see below). If the morphological requirements are satisfied but a connection still cannot be established, syntactic word-class memberships must be involved in the agreement. If the class memberships of both occurrences are relevant and one belongs to a single relevant class, the grammar-code symbol of the other can be changed in the glossary; the same is true if only one occurrence in the pair must belong to a special category. On the other hand, if both occurrences must belong to particular classes and neither belongs to a relevant class, the glossary entries can be

changed only if the classes are unique; otherwise, the pair must be set aside for further analysis. For example, suppose that the connection is possible if the governor belongs to class A and the dependent to class B, or if the governor belongs to class C and the dependent to class D. It follows that the governor belongs to one of two classes, A or C, and the dependent to class B or D, but the information provided by one occurrence is inadequate to make a definite assignment. Other occurrences can make the choice unique if the linguist assumes that the minimum number of assignments per form is desirable, or he can make the decision for each pair.

The possibility of altering grammar-code symbols during SSD raises further problems that must be recognized in the research procedure. The purpose of alteration, roughly speaking, is to prevent connections that are impossible in a certain context. First, a certain word may be restricted to a given class of governors when it is accompanied by one or more dependents of particular types; for example, the object of a preposition sometimes restricts the range of governors that the prepositional phrase can serve, and a genitive singular noun can serve as the subject of a plural verb only if it is accompanied by a cardinal number. Second, the various dependents of a single governor may impose restrictions on one another; most verbs can take a direct object in the genitive case only if they are modified by a negative particle, and a verb cannot take two direct objects. When an entry is added to the table of dependency types, as described above, or a grammar-code symbol is changed in the glossary, the possibility of altering a symbol during SSD is not considered. A screening process can be used thereafter.

In deciding whether alteration of a grammar-code symbol is desirable, negative evidence is needed. The evidence is that a connection between two occurrences is allowed by the dependency table but not by posteditors. The false connection can be eliminated in several ways: by semantic procedures, by subclassification of grammatical categories, or by recognition of contextual restrictions. Only the last leads to alteration of grammar-code symbols. If two grammatical categories are connected by the dependency table, sometimes correctly and sometimes not, a test for contextual restriction should be performed on the pair. As indicated above, there are two cases.

The restriction can involve a chain of three connected occurrences. If the data show that an occurrence of type A governs one of type B only when the latter governs an occurrence of type C, then type B should be altered to type B' when type C is attached, and a dependency-table entry linking types A and B' should replace the AB entry. Type C can be a morphological category, a syntactic word class already established on the basis of other evidence, or a new category established ad hoc, provided that the existence of a class can be shown by the usual statistical evidence of lexical limitation—the number of different words in the class must be less than the number expected by chance.

The restriction can also involve a governor and two of its dependents. If the data show that an occurrence of type A governs one of type B only when it also governs one of type C, then A should be altered to A' when the first of the two dependents is added. Suppose that the AB connection is always earlier than the AC connection in recognition order; then A becomes A' when B is attached, and $A'C$ replaces AC in the dependency table. Type C must satisfy the requirements stated in the preceding paragraph.

The programs described above lead to the establishment of many independent syntactic word classes. Economy demands that the number of distinct classes in the grammar be reduced as much as possible, and it has been suggested that a category is grammatical only if it appears in a number of different rules.[34] The methods and statistical problems of class comparison have been discussed in the section on semantics; the same methods can be applied to syntactic classes, and the statistical problems have to be solved.

One answer to the question of what distinguishes syntactic classes from semantic seems more acceptable than the others. Starting with the notions of morphological classification, function words, occurrence order, and punctuation, the research procedures that have been described here produce certain categories of words. All the word classes that can be defined by rules involving them *and* the initial syntactic indicators are taken as syntactic classes; any class that can be defined by rules involving it and a syntactic class is also a syntactic class. The rules are those that differentiate between structures acceptable to editors and structures that editors reject. It is an empirical question whether a program capable of determining a single acceptable structure for almost every sentence in a large corpus—and more structures than one for almost all of the remainder—can be based entirely on syntactic classes, morphological classes, function words, occurrence order, and punctuation. If the answer is affirmative, then semantics

(by this analysis) is not required for sentence-structure determination; but if the answer is negative, semantics is required for the elimination of *syntactic* ambiguity.

The possibility of writing a grammar for a language by purely automatic methods, using unedited text as data and an analytic program based on linguistic universals, is currently being raised.[35] Although it is still too early to say what results can be obtained with such methods, an important theoretical difference between methods with and without informant editors should be noted at once and remembered as research progresses.

We have seen three levels, or strata,[36] in language: the level of the writing system, the level of the grammar, and the level of semantics. It is apparently characteristic of editor informants—of all users of language—that they deal with all its levels simultaneously and, for the most part, unconsciously. When an informant is asked whether two sound sequences are "same" or "different," he evidently answers according to the grammatical-level patterns that they indicate; as sounds, the sequences can be quite distinct, yet if they stand for the same string of inflected forms, they are "same" to the informant.[37] Two sentences with different composition at the grammatical level are "same" if they are semantically identical, that is, if they indicate the same semantic content; but consciousness reaches the grammatical level, and it is more difficult to apply the test. The point is that informants use their higher-level understanding of a sentence whenever they are asked to comment on it.

An automatic system for grammatical analysis is usually conceived as working its way upward from level to level. First morphological analysis is carried out in accordance with morphological criteria (and lower-level criteria as well; similarity of sound or spelling is used in deciding whether two forms are forms of the same word). Next syntactic analysis is carried out, using morphological and syntactic criteria. Then semantic analysis, using semantic and syntactic criteria, is performed. How far the sequence of levels continues is still an open question, but the proposed automatic analysis programs pass from level to level in one direction only.

If informants and automatic analysis programs operate in exactly opposite directions, are they not certain to yield vastly different results? Perhaps not, for two reasons. A minor point is that informants use criteria at all levels simultaneously; they are not unidirectional. A major point is that language seems universally to have correlated structures on its various levels. The grammatical structure obtained by grammatical criteria corresponds closely with the grammatical structure obtained by semantic criteria. Were this untrue of any language, it would be unspeakably complicated, too complicated for the human organism to learn quickly and use fluently—and if it were learned nevertheless, it would in time be altered for the convenience of its users. Although formal tests of level-to-level structural similarity have never been conducted on a grand scale, the weight of years of linguistic research favors the hypothesis.

Similarity does not imply identity. The syntactically most elegant morphology of a language is not likely to be achieved by following morphological criteria exclusively. The ultimate program for automatic research in linguistics is therefore likely to go forward, then back: A fairly good morphological analysis, based on morphological criteria, paves the way for syntactic analysis; once completed, the syntactic analysis furnishes criteria for adjustment of the morphology. The syntax obtained by using syntactic criteria likewise furnishes the basis for semantic analysis, but the semantic structure, when known, permits refinement of the syntax.[38]

Linguistic methodology is being developed very rapidly; the sound work of recent decades is being tested and enriched by linguists concerned with computers. The criticism sometimes voiced,[39] that computational methods lead to *ad hoc* schemes unthinkingly propounded and not to understanding of the true structure of language, can be refuted if not silenced by attention to some general principles. First, the temptation to overgeneralize must be denied. The modest samples currently available for computational research permit no general statements about languages, and it may be some years before adequate samples can be obtained and analyzed. Second, the search for linguistic universals must continue. Those that are well supported by evidence and relevant to the research now being conducted with computer aid are (1) natural languages can be closely approximated by simple formal models; (2) the appropriate models have recursive features; (3) the appropriate models are multilevel; (4) the appropriate models include simple postulates about occurrence order (at least with respect to separation); (5) the appropriate models include classification of recurrent units (e.g., word classes); (6) the classifications are multidimensional; and (7) simplicity and economy are significant criteria in classification as in the structural design of the model. Third, results obtained by various methods

of research should not propose to refute results obtained by other methods until a more complete, integrated theory of linguistic research is written.

Notes

1. See the discussion beginning with the words "I am concerned with regularities: I am not concerned with rules," in Paul Ziff, *Semantic Analysis*, Cornell University Press, Ithaca, NY, 1960, pp. 34–38.

2. For a broad discussion of the problems involved in evaluating MT output, see George A. Miller and J. G. Beebe-Center, "Some Psychological Methods for Evaluating the Quality of Translations," *Mechanical Translation*, vol. 3, no. 3, pp. 73–80, December, 1956.

3. An early sketch of an MT system was presented by Victor H. Yngve, "A Framework for Syntactic Translation," *Mechanical Translation*, vol. 4, no. 3, pp. 59–65, December, 1957.

4. For example, H. Poutsma, *A Grammar of Late Modern English*, 2d ed., P. Noordhoff, Groningen, 1928 (2 parts in 5 vols.).

5. Zellig S. Harris, *Methods in Structural Linguistics*, University of Chicago Press, Chicago, 1951.

6. The methodological and technical problems raised by the use of informants are enormous. In psychological and sociological research, a sizable literature has grown up. See, for example, Robert L. Kahn and Charles F. Cannell, *The Dynamics of Interviewing*, Wiley, New York, 1957; Herbert Hyman, *Survey Design and Analysis*, The Free Press, Glencoe, Ill., 1955; Warren S. Torgerson, *Theory and Methods of Scaling*, Wiley, New York, 1958.

7. Noam Chomsky, "Semantic Considerations in Grammar," *Georgetown University Monograph Series in Languages and Linguistics*, no. 8, pp. 141–154, Washington, D.C., 1955.

8. Two papers on this subject have recently been published: Paul L. Garvin, "Automatic Linguistic Analysis: A. Heuristic Problem," and Sydney M. Lamb, "On the Mechanization of Syntactic Analysis," in *Proceedings of the International Conference on Machine Translation of Languages and Applied Language Analysis*, vol. 2, pp. 655–686, H. M. Stationery Office, London, 1962. See also O. S. Kulagina, "A Method of Defining Grammatical Concepts on the Basis of Set Theory," *Problemy Kibernetiki*, no. 1, pp. 203–214, 1958.

9. A point raised by I. I. Revzin, "On the Notion of a 'Set of Marked Sentences' in the Set-theoretic Concept of O. S. Kulagina," in N. D. Andreyev (ed.), *Abstracts of the Conference on Mathematical Linguistics*, Leningrad, 1959. Translation 893-D, U.S. Joint Publications Research Service, Washington, D.C., 1959.

10. The latest edition of the guide used in this work at RAND is Kenneth E. Harper et al., Studies in Machine Translation, 8: *Manual for Postediting Russian Text*, RM-2068, The RAND Corporation, Santa Monica, Calif., 1960.

11. Louis Hjelmslev, *Prolegomena to a Theory of Language* (Francis J. Whitfield, tr.), Univ. of Wisconsin Press, Madison, Wis., 1961, pp. 16–18.

12. L. R. Micklesen, "Russian-English MT," in Erwin Reifler (ed.), *Linguistic and Engineering Studies in Automatic Language Translation of Scientific Russian into English*, Univ. of Washington Press, Seattle, Wash., 1958, p. 5.

13. Lucien Tesnière, *Eléments de Syntaxe Structurale*, Klincksieck, Paris, 1959.

14. Y. Lecerf, "Programme des Conflits, Modèle des Conflicts," *La Traduction Automatique*, vol. 1, no. 4, pp. 11–20, October, 1960, and vol. 1, no. 5, pp. 17–36, December, 1960.

15. Kenneth E. Harper and David G. Hays, "The Use of Machines in the Construction of a Grammar and Computer Program for Structural Analysis," *Information Processing*, UNESCO, Paris, 1960, pp. 188–194. David G. Hays, "Grouping and Dependency Theories," in H. P. Edmundson (ed.), *Proceedings of the National Symposium on Machine Translation*, Prentice-Hall, Englewood Cliffs, N.J., 1961, pp. 258–266. Haim Gaifman, *Dependency Systems and Phrase Structure Systems*, P–2315, The RAND Corporation, Santa Monica, Calif., 1961.

16. See, for example, K. E. Harper, D. G. Hays, and D. V. Mohr, *Studies in Machine Translation—6: Manual for Coding Russian Grammar*, RM-2066-1, The RAND Corporation, Santa Monica, Calif., 1958 (rev. 1960). A. S. Kozak et al., *Studies in Machine Translation—12: A Glossary of Russian Physics*, RM-2655, The RAND Corporation, Santa Monica, Calif., 1961.

17. This is only a brief statement of grammatical theory; for a more complete treatment see Charles F. Hockett, *A Course in Modern Linguistics*, Macmillan, New York, 1958.

18. I. A. Mel'chuk, "Machine Translation and Linguistics," in O. S. Akhamanova et al., *Exact Methods in Linguistic Research*, Moscow University Press, Moscow, 1961.

19. Abraham Kaplan, "An Experimental Study of Ambiguity," *Mechanical Translation*, vol. 2, no. 2, pp. 39–46, November, 1955.

20. Zellig S. Harris, "Distributional Structure," *Word*, vol. 10, pp. 146–162, 1954.

21. The author is indebted to Duane G. Metzger and A. Kimball Romney for the point of view adopted here.

22. The author is indebted to Martin J. Kay for discussions of this point.

23. See, for example, Kenneth E. Harper, "Procedures for the Determination of Distributional Classes," in *Proceedings of the International Conference on Machine Translation of Languages and Applied Language Analysis op. cit.*, fn. 8, vol. 2, pp. 699–700. [. . .]

31. Margaret Masterman, "Potentialities of a Mechanical Thesaurus," *Mechanical Translation*, vol. 3, no. 2, p. 36, November, 1956. See also her paper in *Proceedings of the International Conference on Machine Translation and Applied Linguistic Analysis*, *op. cit.*, fn. 8, vol. 2, pp. 437–474.

33. Kenneth E. Harper, *Machine Translation of Russian Prepositions*, Paper P1941, The RAND Corporation, Santa Monica, Calif., 1960.

34. Edward Klima, personal communication.

35. See references cited in fn. 8.

36. Sydney M. Lamb, "The Strata of Linguistic Structure," presented at a meeting of the Linguistic Society of America, Hartford,

Conn., December, 1960. (His strata are not identical with these.) *Cf.* also the three sets of levels in Garvin, "The Definitional Model of Language."

37. See Chomsky, *op. cit.*, fn. 7.

38. David G. Hays, "Linguistic Methodology and the Theory of Strata," presented at a meeting of the Philological Association of the Pacific Coast, Santa Barbara, Calif., November, 1961.

39. Dean S. Worth, "Linear Contexts, Linguistics, and Machine Translation," *Word*, vol. 15, no. 1, pp. 183–191, April, 1959. Noted with agreement and expanded by Mel'chuk, *op. cit.*, fn. 18.

This page intentionally left blank

11
ALPAC: The (In)Famous Report

John Hutchins

The best known event in the history of machine translation is without doubt the publication thirty years ago in November 1966 of the report by the Automatic Language Processing Advisory Committee (ALPAC 1966). Its effect was to bring to an end the substantial funding of MT research in the United States for some twenty years. More significantly, perhaps, was the clear message to the general public and the rest of the scientific community that MT was hopeless. For years afterwards, an interest in MT was something to keep quiet about; it was almost shameful. To this day, the 'failure' of MT is still repeated by many as an indisputable fact.

The impact of ALPAC is undeniable. Such was the notoriety of its report that from time to time in the next decades researchers would discuss among themselves whether "another ALPAC" might not be inflicted upon MT. At the 1984 ACL conference, for example, Margaret King (1984) introduced a panel session devoted to considering this very possibility. A few years later, the Japanese produced a report (JEIDA 1989) surveying the current situation in their country under the title: "A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC."

While the fame or notoriety of ALPAC is familiar, what the report actually said is now becoming less familiar and often forgotten or misunderstood—and this extensive summary includes therefore substantial extracts.

The report itself is brief—a mere 34 pages—but it is supported by twenty appendices totalling a further 90 pages. Some of these appendices have had an impact as great as the report itself, in particular the evaluation study by John Carroll in Appendix 10.

The first point to note is that the report is entitled: "Languages and Machines: Computers in Translation and Linguistics." It was supposedly concerned, therefore, not just with MT but with the broader field of computational linguistics. In practice, most funded NLP research at the time was devoted to full-scale MT.

The background to the committee is outlined in the Preface:

> The Department of Defense, the National Science Foundation, and the Central Intelligence Agency have supported projects in the automatic processing of foreign languages for about a decade; these have been primarily projects in mechanical translation. In order to provide for a coordinated federal program of research and development in this area, these three agencies established the Joint Automatic Language Processing Group (JALPG).

It was the JALPG which set up ALPAC in April 1964 under the chairmanship of John R. Pierce (at the time, of Bell Telephone Laboratories). Other members of the committee were John B. Carroll (Harvard University), Eric P. Hamp (University of Chicago), David G. Hays (RAND Corporation), Charles F. Hockett (Cornell University, but only briefly until December 1964), Anthony G. Oettinger (Harvard University), and Alan Perlis (Carnegie Institute of Technology). Hays and Oettinger had been MT researchers, although no longer active when ALPAC was meeting (having become disillusioned with progress in recent years); Perlis was a researcher in artificial intelligence; Hamp and Hockett were linguists; and Carroll was a psychologist. The committee did, however, hear evidence from active MT researchers such as Paul Garvin and Jules Mersel (Bunker–Ramo Corporation), Gilbert King (Itek Corporation and previously IBM), and Winfred P. Lehmann (University of Texas).

The committee agreed at the outset that support for research in this area "could be justified on one of two bases: (1) research in an intellectually challenging field that is broadly relevant to the mission of the supporting agency and (2) research and development with a clear promise of effecting early cost reductions, or substantially improving performance, or meeting an operational need." ALPAC rejected (1), deciding that the motivation for MT research was the practical one of (2) alone. For this reason, ALPAC "studied the whole translation problem" and whether MT had a role in it.

The second point to note, therefore, is that the report concentrated exclusively on US government and military needs in the analysis and scanning of Russian-language documents. It was not concerned in any way with other potential uses or users of MT systems or with any other languages.

The first half of the report (pp. 1–18) investigated the translation needs of US scientists and government officials and overall demand and supply of translations from Russian into English. ALPAC began by asking whether, with the overwhelming predominance of English as the language of scientific literature (76% of all articles in 1965), it "might be simpler and more economical for heavy users of Russian translations to learn to read the documents in the original language." Studies indicated that this could be achieved in 200 hours or less, and "an increasing fraction of American scientists and engineers have such a knowledge," and it noted that many of the available opportunities for instruction were underutilized (appendix 2).

Next it looked at the supply of translations within government agencies (including those sponsoring MT research). They used a combination of contract and in-house translators. The committee was not able to determine the exact number of in-house translators, but it did establish that the average salary of translators was markedly lower than that of government scientists. Nevertheless, it found "a very low rate of turnover among government translators. Indeed, the facts are that the supply exceeds demand." At the time of the report, no post of government translator was vacant while there were over 500 translators registered in the Washington area (statistics in appendix 8 of the report).

The committee was thus prompted to ask whether there was any shortage of translators. The Joint Publications Research Service, it found, had the capacity to double translation output immediately: out of 4000 translators under contract only 300 on average were being used each month. Likewise, the National Science Foundation's Publication Support Program was prepared to support the cover-to-cover translation of any journal which might be nominated for complete translation by any "responsible" society. Appendix 6 recorded 30 journals being translated from Russian in this way during 1964. Since some had very low circulations (appendix 6), ALPAC questioned the justification for this virtually "individual service."

Indeed, ALPAC wondered whether there were not perhaps an excess of translation, on the argument that "translation of material for which there is no definite prospective reader is not only wasteful, but it clogs the channels of translation and information flow." What it found was that many Russian articles were being translated which did not warrant the effort: according to a 1962 evaluation, only some 20 to 30% of Russian articles in some fields would have been accepted for publication in American journals; furthermore the delays in publication of cover-to-cover translations reduced their value. The committee concluded that the main need was for "speed, quality, and economy in supplying such translations as are requested."

At this point, before considering MT as such, the report anticipated its conclusions with the bold statement (p. 16): "There is no emergency in the field of translation. The problem is not to meet some nonexistent need through nonexistent machine translation. There are, however, several crucial problems of translation. These are quality, speed, and cost."

On quality, ALPAC stressed that it must be appropriate for the needs of requesters: "flawless and polished translation for a user-limited readership is wasteful of both time and money." But there were no reliable means of measuring quality, and for this reason ALPAC set up an evaluation experiment (reported in appendix 10). This study by John B. Carroll evaluated both human and machine translations, and it had great influence on many MT evaluations in subsequent years. It was supplemented in appendix 11 by a study from the Arthur D. Little, Inc. of MT errors, based on the system in use at the time at the Foreign Technology Division, i.e., the system developed by Gilbert King at IBM.

On speed, ALPAC saw much room for improvement: scientists were complaining of delays; the most rapid service (from JPRS) was 15 days for 50 pages; the NSF translation of journals ranged from 15 to 26 weeks; documents sent to outside contractors by the US Foreign Technology Division were taking a minimum of 65 days; and when processed by the FTD's MT system, they were taking 109 days (primarily caused by processes of postediting and production, detailed in appendix 5).

On cost, ALPAC considered what government agencies were paying to human translators and this varied from $9 to $66 per 1000 words. In appendix 9 calculations were made of cost per reader of the different forms of translation, including unedited output from the FTD system. These costs included the expenditure of time by readers. Assuming that the average reader took twice as long to read unedited MT documents as good quality human translation (based on the results of Carroll's evaluation in appendix 10), it concluded that if documents are to be

read by more than 20 persons traditional human translation was cheaper than MT. As for the costs of postedited MT, they would include posteditors proficient in Russian; ALPAC concluded that "one might as well hire a few more translators and have the translations done by humans ... [or] take part of the money spent on MT and use it either (1) to raise salaries in order to hire bilingual analysts—or, (2) to use the money to teach the analysts Russian."

At this point, the report turned to "the present state of machine translation" (pp. 19–24). It began with a definition: MT "presumably means going by algorithm from machine-readable source text to useful target text, without recourse to human translation or editing." And immediately concluded: "In this context, there has been no machine translation of general scientific text, and none is in immediate prospect."

Support for this contention, ALPAC asserted, came from "the fact that when, after eight years of work, the Georgetown University MT project tried to produce useful output in 1962, they had to resort to postediting. The postedited translation took slightly longer to do and was more expensive than conventional human translation." Likewise, ALPAC regarded it as a failure that the MT facility at FTD "postedits the machine output when it produces translations."

However, the principal basis for its conclusion was the results of Carroll's evaluation exercise in appendix 10. "Unedited machine output from scientific text is decipherable for the most part, but it is sometimes misleading and sometimes wrong ... and it makes slow and painful reading." The report then printed (pp. 20–23) what it held to be "typical" samples of the "recent (since November 1964) output of four different MT systems." These were presumably those used in the evaluation exercise, but this was not stated explicitly. The four systems were from Bunker–Ramo Corporation, from Computer Concepts, Inc., from the USAF Foreign Technology Division, and from EURATOM. The first would have been the system developed by Paul Garvin after he left Georgetown in 1960. The EURATOM system was the Georgetown University system installed in 1963 at Ispra, Italy. The FTD system was, as already mentioned, the one developed by Gilbert King at IBM, using his patented photoscopic store (a precursor of the laser disk). The Computer Concepts company had been set up by Peter Toma after he left the Georgetown project in 1962; the system illustrated was presumably AUTOTRAN, based in many respects on the SERNA version of the Georgetown system, and a precursor of SYSTRAN. Only the EURATOM and FTD systems were fully operational at this time, the other two were still experimental prototypes—but this was not mentioned by ALPAC.

After reproducing the MT samples, the report continued: "The reader will find it instructive to compare the samples above with the results obtained on simple, selected, text 10 years earlier (the Georgetown IBM Experiment, January 7, 1954) in that the earlier samples are more readable than the later ones." Twelve sentences from the highly-restricted demonstration model (Hutchins 1994) are then listed, with the comment: "Early machine translations of simple or selected text ... were as deceptively encouraging as 'machine translations' of general scientific text have been uniformly discouraging."

There can be no doubt about the deficiencies and inadequacies of the translations illustrated but it was perhaps a major flaw of ALPAC's methodology to compare unfavorably the results of general-purpose MT systems (some still experimental) working from unprepared input (i.e. with no dictionary updating) and the output of a small-scale demonstration system built exclusively to handle and produce a restricted set of sentences.

ALPAC concluded this chapter by stating that it was very unlikely that "we will suddenly or at least quickly attain machine translation," and it quoted Victor Yngve, head of the MT project at MIT, that MT "serves no useful purpose without postediting, and that with postediting the overall process is slow and probably uneconomical." However, the committee agreed that research should continue "in the name of science, but that the motive for doing so cannot sensibly be any foreseeable improvement in practical translation. Perhaps our attitude might be different if there were some pressing need for machine translation, but we find none."

At this point, ALPAC looked at what it considered the much better prospects of "machine-aided translation" (not, as it stressed, human-aided MT, but what are now referred to as translation tools). It had high praise for the production of text-related glossaries at the Federal Armed Forces Translation Agency in Mannheim (Germany) and for the terminological database at the European Coal and Steel Community, which included terms in sentence contexts—this was the precursor of EURODICAUTOM. (Further details were given in appendices 12 and 13, pp. 79–90). Its general conclusion was that these aids, primitive as they were, were much more economically effective in the support of translation than any MT systems.

The alternative it saw was postedited MT. However, it admitted that it could not "assess the difficulty

and cost of postediting.'' Appendix 14 (pp. 91–101) reported on a study involving the translation of two excerpts from a Russian book on cybernetics, and the postediting of an MT version of one of the excerpts. Interestingly, ''eight translators found postediting to be more difficult than ordinary translation. Six found it to be about the same, and eight found it easier.'' Most translators ''found postediting tedious and even frustrating,'' but many found ''the output served as an aid ... particularly with regard to technical terms.'' Despite the inconclusiveness of this study, ALPAC decided to emphasise the negative aspects in the body of its report, quoting at length the comments of one translator:

> I found that I spent at least as much time in editing as if I had carried out the entire translation from the start. Even at that, I doubt if the edited translation reads as smoothly as one which I would have started from scratch. I drew the conclusion that the machine today translates from a foreign language to a form of broken English somewhat comparable to pidgin English. But it then remains for the reader to learn this patois in order to understand what the Russian actually wrote. Learning Russian would not be much more difficult.

At the beginning of the next chapter ''Automatic Language Processing and Computational Linguistics,'' ALPAC made one of its most often cited statements, namely that ''over the past 10 years the government has spent, through various agencies, some $20 million on machine translation and closely related subjects.'' The statistics provided in appendix 16 (pp. 107–112) reveal that by no means all this sum was spent on MT research in the United States. Firstly, the total includes $35,033 on sponsoring three conferences and $59,000 on ALPAC itself. Secondly, it includes $101,250 in support of research outside the United States (at the Cambridge Language Research Unit) and $1,362,200 in support of research under Zellig Harris at the University of Pennsylvania which even at the time was not considered to be directly related to MT. Thirdly, it lists global sums from the US Air Force, US Navy and US Army (totalling $11,906,600) with no details of the recipients of the grants. Evidence from elsewhere (details in Hutchins 1986, p. 168) suggests that much of the funding was in support of developments in computer equipment rather than MT research (perhaps up to two thirds of the USAF grants). In brief, the funding of US agencies on US research in MT may well have been nearer $12–13 million than the frequently repeated $20 million stated by ALPAC. The sum was still large, of course, and ALPAC was right to emphasise the poor return for the investment.

The main theme of this chapter on ''Automatic Language Processing and Computational Linguistics'' was a consideration of the contribution of MT research to advances of NLP in general. Summarizing the more extensive findings in appendices 18 and 19, it found that its effect on computer hardware had been insignificant, that it had contributed to advances in ''computer software (programming techniques and systems),'' but that ''by far the most important outcome ... has been its effect on linguistics.'' Here they highlighted insights into syntax and formal grammar, the bringing of ''subtler theories into confrontation with richer bodies of data,'' and concluding that although ''the revolution in linguistics has not been solely the result of attempts at machine translation and parsing ... it is unlikely that the revolution would have been extensive or significant without these attempts.'' (This is a view which would certainly be disputed today.) However, despite this favourable influence, ALPAC did not conclude that MT research as such should continue to receive support; rather it felt that what was required was

> basic developmental research in computer methods for handling language, as tools for the linguistic scientist to use as a help to discover and state his generalizations, and ... to state in detail the complex kinds of theories ..., so that the theories can be checked in detail.

In the final chapter (pp. 32–33), ALPAC underlined once more that ''we do not have useful machine translation [and] there is no immediate or predictable prospect of useful machine translation.'' It repeated the potential opportunities to improve translation quality, particularly in various machine aids: ''Machine-aided translation may be an important avenue toward better, quicker, and cheaper translation.'' But ALPAC did not recommend basic research: ''What machine-aided translation needs most is good engineering.''

ALPAC's final recommendations (p. 34) were, therefore, that research should be supported on:

> 1. practical methods for evaluation of translations; 2. means for speeding up the human translation process; 3. evaluation of quality and cost of various sources of translations; 4. investigation of the utilization of translations, to guard against production of translations that are never read; 5. study of delays in the over-all translation process, and means for eliminating them, both in journals and in individual items; 6. evaluation of the relative speed and cost of various sorts of machine-aided translation; 7. adaptation of existing mechanized editing and production processes in translation; 8. the over-all translation process; and 9. production of adequate reference works for the trans-

lator, including the adaptation of glossaries that now exist primarily for automatic dictionary look-up in machine translation.

Aware that these recommendations failed to support not just MT but any kind of natural language processing, a statement was inserted in the final report addressed to the president of the National Academy of Sciences from the chairman John R. Pierce in which he stressed the value of supporting "computational linguistics, as distinct from automatic language translation." Elaborating on recommendations in its chapter on NLP, the chairman believed that the National Science Foundation should provide funds for research on a reasonably large scale, "since small-scale experiments and work with miniature models of language have proved seriously deceptive in the past,"—obviously alluding to MT experience—"and one can come to grips with real problems only above a certain scale of grammar size, dictionary size, and available corpus."

The ALPAC report was relatively brief; and its direct discussion of MT amounted to just one chapter (pp. 19–24) and four appendices (on evaluating translation (pp. 67–75), on errors in MT (pp. 76–78), on postediting MT compared with human translation (pp. 91–101), and on the level of government expenditure on MT (pp. 107–112). The rest of the report was concerned with the demand for translation in general by US government agencies, the supply of translators, with computer aids for translators, and with the impact of MT on linguistics. However, it was in these few pages that ALPAC condemned MT to ten years of neglect in the United States (longer, as far as government financial support was concerned), and it left the general public and the scientific community (particularly researchers in linguistics and computer science) with the firm conviction that MT had been a failure or, at best, very unlikely to be a useful technology.

In some respects, the impact of ALPAC can be exaggerated. MT research in the US did not come to a complete and sudden halt in 1966. Some projects continued, notably at Wayne State University under Harry Josselson until 1972 and at the University of Texas under Winfred Lehmann and Rolf Stachowitz until 1975 (later revived in 1978 with funding from Siemens). Furthermore, some MT projects supported by government money had ended before ALPAC reported: University of Washington (1962), University of Michigan (1962), Harvard University (1964). In particular, the Georgetown University project, whose system was explicitly criticized by ALPAC, had received no funding after 1963. By this time it had installed operational MT systems at the Oak Ridge National Laboratory and at the EURATOM laboratories in Italy.

Furthermore, in hindsight it can, of course, be agreed that ALPAC was quite right to be sceptical about MT: the quality was undoubtedly poor, and did not appear to justify the level of financial support it had been receiving. It was also correct to identify the need to develop machine aids for translators, and to emphasize the need for more basic research in computational linguistics. However, it can be faulted for concentrating too exclusively on the translation needs of US scientists and of US agencies and not recognizing the broader needs of commerce and industry in an already expanding global economy. In this way, ALPAC reinforced an Anglo-centric insularity in US research which damaged that country's activities in multilingual NLP at a time when progress continued to take place in Europe and Japan. It took two decades for the position to begin to be rectified in government circles, with the report for the Japan Technology Evaluation Center (JTEC 1992) and with ARPA support of US research in this field during the 1990s.

References

ALPAC (1966) *Languages and machines: computers in translation and linguistics*. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council. Washington, D.C.: National Academy of Sciences, National Research Council, 1966. (Publication 1416.)

Hutchins, W. J. (1986) *Machine translation: past, present, future*. Chichester: Ellis Horwood.

Hutchins, W. J. (1994) The Georgetown-IBM demonstration, 7th January 1954, *MT News International*, no. 8 (May 1994), 15–18.

JTEC (1992) *JTEC Panel report on machine translation in Japan*. Jaime Carbonell et al., Baltimore, MD: Japanese Technology Evaluation Center, Loyola College in Maryland, January 1992.

JEIDA (1989) *A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, U.S.A.* Tokyo: JEIDA, 1989.

King, M. (1984) When is the next ALPAC report due? In: *10th international conference on computational linguistics. Proceedings of Coling '84*, July 1984, Stanford University, CA (ACL, 1984), pp. 352–353.

This page intentionally left blank

12
Correlational Analysis and Mechanical Translation

Silvio Ceccato

The Three Approaches to Mechanical Translation

By now we know that there is more than one way to attack the problem of mechanical translation. I know of at least three which have already been followed, and I can imagine at least a fourth one.

First we must isolate a way of approaching mechanical translation which avoids all linguistics. We can quickly see how this might happen, by remembering what a bilingual dictionary is. In this case a man who knows two languages presents the translation of part of the words of the two languages, that is, he shows how those of the first are substituted for those of the second. A little grammar allows the person who uses the dictionary to obtain those words from all the words in the language (for example, by reducing a plural to a singular, a tense, mood, and person of a verb to the infinitive, etc.), and to obtain all the words in the language from those words. But, if the person who compiles the dictionary takes care of it himself, even that grammar can be eliminated. The bilingual dictionary will then contain all the desired substitutions for single words; the machine will do nothing but register them in its memory. Continuing in this same direction it is possible that the compiler further enriches his dictionary, making it contain a certain number of already translated expressions. Thus would be obtained a machine which "translates" without applying any grammatical and logical analysis, and therefore without the builder having had to resolve any problem of linguistics, philosophy or psychology.

In such a way, the first practical results of mechanical translation were obtained. But we understand immediately the limitations of this system and how these limits are not even surmountable. These derive from the fact that some words taken by themselves have more than one meaning [...] which is made clear only in the context [...]; furthermore, only sometimes do two languages correspond exactly in the composition of things to designate [...], with the consequence that, granted the impossibility of finding the term which corresponds to a single word, it will be necessary to try to find approximate equivalents in whole expressions. On the other hand, it is impossible to introduce all possible sentences, already translated, into the machine. [...]

Because of this, it was to be expected that [...] even those who had departed from the principle of mechanizing a bilingual dictionary would feel the need of enriching the program of the translating machine with classifications and rules which would permit the machine to define clearly the meaning of words. That is, it was decided to add grammar, both monolingual and bilingual.

At this point the entrepreneurs of mechanical translation must have been unpleasantly surprised for grammar, as it was conceived for men, is not immediately applicable to machines. It was born not so much as a key for interpreting discourse, and thus to pass from the words to the designated thought; but rather to systematize language, with only passing reference to thought. [...]

If the grammar which serves men is not immediately applicable to machines, how must it be enlarged, how completed? [...] We understand immediately that a grammar, conceived not for men but for a machine, must be studied keeping in mind that in it which is due to thought, to facts, etc. Thus we understand how such a grammar encounters the classical problems that philosophy and psychology, and, in fact, nearly all disciplines have encountered for so long. It is true that in translation one starts not from a situation of thought or reality to be designated, but rather from a situation which is already linguistic: but an illusion is hidden in this consideration, because what one starts from [...] is already language in that it refers to something else. What this something else is remains the problem which must be formulated and resolved.

Aside from the scholastic grammars, might there not exist a linguistic scheme among the so-called "reasoned" linguistics, in which language is studied with its counterpart of thought or reality? Attempts

have been made in this direction (Tesnière [1959], Brøndal [1943], Hjelmslev [1935]) [...]. However the results actually produced are far from adequate for a machine to master language. To understand this insufficiency it is enough to consider the exigencies which a description must satisfy so that it can be employed on a machine. In short, everything must be presented in terms of operations [...].

Now, if we open the book by Tesnière, we will find such statements as these from the very beginning: "La phrase est un *ensemble organisé* dont les éléments constituants sont les mots"; "Entre (un mot) et ses voisins, l'esprit aperçoit des *connexions* ..."; "Ces connexions ne sont indiquées par rien. Mais il est indispensable qu'elles soient aperçues par l'esprit ..." (p. 11). But it is clear to any builder of machines that even if he understands what Tesnière means, Tesnière's statements are not of any help to him. What does "ensemble organisé" mean? Nothing, if a declaration of the criterion of organization is lacking. For example, an "ensemble" of dishes could be organized in all sorts of ways: on a sideboard, waiting to be used at dinner; arranged on a table; in an exhibit; etc. As to the connections between words, indicated with "rien" and "aperçues par l'esprit", the sentence will certainly seem magical and paralyzing as soon as an engineer tries to mechanize this subject. And yet, those who have read Tesnière's book and have had the patience to follow us in this paper can verify whether what he says is tenable and can be presented in technical terms up to a certain point. It was necessary then to construct a new linguistics. The questions: what is grammar? what is language? etc., could find answers in this linguistics suitable both for men interested in a dynamic operative linguistics, and for builders of linguistic machines. A description of things in terms of operations is in fact typically mechanical [...].

In 1955, I had been trying for some time to reach a description of thought, of language, and of their relationship without metaphorical or negative expressions. Even before thinking in mechanical terms, I had come to the conclusion (Ceccato [1960, 1961, 1963, 1965a,b]) that a fertile direction of investigation for linguistics, philosophy and psychology lay in giving a description in terms of operations of various named things, whether these were named by single words or by various possible successions of words. My attempt was successful, at least in part; and I tried to apply the results I had obtained to MT.

This stage in the research naturally supposed that various traditional conceptions had been overcome, for example, that of words which designate and those which do not. [...] I shall now report briefly on this work, on the prospects which it offers for MT, and at the same time on the limitations which it brings to light.

When man is studied in terms of activities, the two ways in which he operates must be kept distinct.

With one activity we modify things: this is the activity which is called manual or physical, the usual activity of the hands, feet, stomach, kidneys, and so forth. It comes into play when we knead bread, when we move our body, when we digest, etc., and it is continued in the foundry, the workshop, etc.

With another activity, however, we modify nothing at all; for example when we perceive things, when we represent them to ourselves or when we categorize them. This is the activity with which we think.

Among the many consequences of this is the fact that, if we can speak of operations and results in connection with both activities, in the modifying activity, results can be reached by more than one path at least in theory (consider the ways of producing sulphuric acid); whereas there is only one path to the results of the constitutive activity (a certain object, in that it is perceived or represented, is composed of certain operations; the singular and the plural are certain operations and not others; etc.).

This distinction is important in the study of the speaking man in terms of activities, because in discourse the two activities are both present, the constitutive one for the thought and all its possible contents, and the modifying one in the modulation of sounds, or the tracing of signs, that is, for producing words. In this connection, note that, although the constitutive activity, like every activity, can be considered a function of organs and therefore ascribed to their functioning, it is never observable in the changes shown by the organs in their functioning. The obstacle is overcome by finding a parallelism between the two activities; and the choice of the modifying activity which produces sounds or marks as the counterpart of the constitutive activity answers, among other things, the fundamental requirement of being easily producible, transferable, depositable, and so on. [...]

We have asserted that in discourse the two types of activity intervene, with connections due to linguistic conventions. This assertion could appear simply as a working hypothesis, if the research already carried out had not shown that in fact every linguistic expression, from the single word to the most complex expression, can be examined in precisely these terms.

In planning a linguistic machine, a "machina loquax" which reproduces human linguistic behavior,

we should thus direct our research in two directions, examining (1) what the units of connection are, that is, the units by means of which the two activities are connected, and how these units are combined, (2) what the operative units are, that is, the simple operations or their combinations, which enter into the constitution of things designated in the units of connection.

As is usually the case when human behavior is studied, there are many differences, which here appear as the "oddities" of the single languages, but there is also a strong uniformity, not only between those who speak the same language, but also among those who speak different languages. It is this uniformity on which the possibility of communication and understanding is based. [...]

Discourse [...] designates operations constitutive of things [...]. But this dynamism is still not sufficient for something to be present to us, in order for us to be aware of it, at least if this thing is not attention itself. [...] We find this attention first of all by itself, pure, as the state we assume when someone says to us, "watch out!", "look!", "listen!" and the like; the attention then is not focused. This attention can be taken to various places, to the inside of our body, to the epidermis, or even outside our body, giving our organs the desired orientation and focus. [...]

The application of the attention, beside bringing to mind the functioning of the other organs, isolates a tantum of it, permitting a unit of designation to be fixed, in a flow—in which otherwise there would be no beginning nor end—which thus is fragmented. For example, the clock may continue to beat; but with the play of our attention, we can isolate a "tick" or a "tock" once, and another time we can encompass an entire "tick-tock", that is, the two noises and the pause.

[A discussion of the combinations of states of attention and the underlying model of discourse is omitted—*Eds.*]

When men communicate with each other, it is useless to recount what the others already know. It is enough to think of the difference in the number of indications which we furnish when we communicate something to somebody who is closely connected with us in life, or in our business, etc., and to someone whom we have met for the first time. In telling someone "Call me tomorrow at home" it would be senseless to tell him precisely where I live and at what hours he can find me, if he already knows it; these indications are useful only for strangers.

Now, there are many things which those who belong to a certain society all know, in that they enter not only in the most elementary widespread culture, but also in their most common experiences of life. For example, everyone knows that certain things are eaten and others are not, that certain things are drunk and others are not, that in cities one finds streets and houses, that books are made of pages and a cover, etc. This situation is mirrored however in discourse, as it appears as a dearth of explicit indications. For example, in "He eats and drinks vodka" everyone understands that "vodka" is the object of "to drink" but not "to eat", precisely because vodka, being liquid, is drunk, and not eaten, even if nothing in the sentence indicates that the transitivity of "to eat" must be stopped before the object of "to drink". If we have written "He buys, sells, and drinks vodka", the three developments might all have had vodka as an object. Cases of this sort are not rare; in some languages, such as Latin or Italian, they are frequent. Position, that is, is significant, but not in a completely univocal way. As we have seen, when a whole correlation becomes an element of a larger correlation it is necessary that the indications which it contains be isolated, by adding to them, considered as a unit, the indication of the function of the unit within the new correlation. This is easy enough for a person speaking, who can distribute accents and pauses, but in writing punctuation is often not sufficient to substitute these indications, notwithstanding the comma, which has, among its functions, that of breaking words into groups. If the dearth of indications does not hinder communication between men, it is because it is supplied by that which everyone knows by other means.

This knowledge has principally two sources. The one is our operating itself, above all our perceptive-representative operating. In representing things to ourselves, we immediately see certain connections, certain compatibilities and incompatibilities; for example, in order to continue the preceding illustration, it is impossible to see someone eating, that is, chewing, a liquid; for this reason, vodka is excluded without any doubt as an object of eating. The other source of knowledge is the enrichment which this operating receives from culture. For example, only the knowledge that "ave" was the usual Roman form of respectful greeting guides us to understand immediately the "ave" in the famous Latin sentence "Ave Caesar, morituri te salutant" as a form of salutation. Otherwise we might happen to translate the "ave" as a young boy told me, as the ablative case of "avis", bird: "Caesar, those who are about to die salute

you with the bird'', a situation which he imagined as a raising of the eagles of the insignia carried as standards.

The dearth of explicit information, if it does not create difficulties for man, but rather assures him an economic and quick discourse, is troublesome both when he wants to find an algorithm which describes language, and when he wants to mechanize our linguistic activity, and in particular our comprehension of language.

We must, in fact, prepare a system of linguistics which distinguishes that which, in the relationship between thought and language, appears explicitly from that which implicitly enters into it; we must also, in realizing this project, keep in mind the present limitations of our constructional technique, besides those of time and money.

In the meantime, I shall make a survey of the present situation.

[...] Any linguistic study conducted on man and for man, presupposes not only [...] thought activity, but also the knowledge of at least one language, the mother tongue. The linguistic machine, instead, does not exist except insofar as it has been constructed; and today we are far from having constructed machines which can perform activities of thought and which lack only the addition of a certain number of semantic connections in order to make them into talking machines.

Even after the analyses which have been conducted on thought and language there are still many areas of shadow, and, in any case, at least one difficulty, which appears to be insuperable. This is the construction of a memory with the characteristics of human memory, which is associative, selective, and propulsive. In describing how the various contents of thought are constituted, with holdings and summarizing revivals, we have already shown that memory is an essential factor. The integration of the specific linguistic indications, through experience of life and culture, reveals its irreplaceable function in the comprehension of language.

But there is more. Whoever follows a speech, or reads a text, carries into each successive sentence that which he had learned in the preceding one. This guides him greatly in comprehension. As the discourse progresses the context of each word broadens, and makes any doubt difficult. Culture, experience of life, and the preceding statements already anticipate so much of what is about to be said that we ourselves are often in a position to continue and to conclude. Linguistic expression, that is, rather than marking out a road for our thought, seems sometimes to limit itself to giving thought a push on a road which is already well marked.

All this is due however precisely to a memory which, like the human one, is not only associative, selective, but also propulsive, that is, it does not need to be recalled, but it pushes forward, it leads us incessantly.

Further, through this dynamic, active memory, and through the culture and experiences which it revives, we add, as units of comprehension of discourse, units already complete enough to exclude any alternative correlational net. Whoever attempts to read a text, uncovering it word by word, will immediately realize that a single word is open to a certain number of possible meanings, above all with regard to its grammatical categorization, and the value, still undetermined, which derives from its position in the succession of words. He will also realize that this way of reading is not at all his normal one, by which he encompasses units of several words which clarify each other's meanings.

Linguistics for machines, at least for the time being, will have to leave to one side that which is permitted to man thanks to the characteristics of his memory. [...]

Linguistics for Machines

Keeping these theoretical and practical limitations in mind, I have laid down the lines of a system of linguistics for machines, conceived for MT, by taking the following decisions:

1. Single words are assumed as input units. They are analyzed according to what they designate, whether of the contents of thought, or of the correlational function of thought contents [...].

2. The contents of thought are designated by the word or parts of a word which designate them, while their constitutive operations are substituted by a certain number of classifications; their correlational functions are indicated with reference to the contents of thought which can act as correlators, because of their position as correlata in relationship to these correlators [...].

3. From the classifications, the contents of thought receive all the combinational possibilities which belong to them, that is, independently from the place in which the word which designates them is found in the

Table 12.1

correlator	
I correlatum	II correlatum

discourse; and then rules are fixed for reducing these possibilities when the word which designates them encounters other words. [...]

4. The context of the encounter of each word with the others is limited to the range of one proposition, and from certain points of view, to that of two.

5. Our culture and experience are substituted through a notional sphere, in which each designatum appears in its most common relationships with each of the other designata. [...]

6. Transformations, if any, to be effected on the correlational net of the input text, so that it can be expressed in the output text, are prepared through constellations in which appear certain relationships subsisting between developments and their subjects, objects and other complements, rather than between the characterized things and their characteristics, by indicating how these relationships are posed and expressed in the two languages.

I shall now illustrate briefly the various points of this system of linguistics. For this purpose it is worth introducing a convenient graphic representation of the correlational structure (see table 12.1), by resorting to three boxes, of which the upper one is reserved for the correlator, the one at the lower left is reserved for the first correlatum, and the one at the lower right for the second correlatum.

Meanwhile, it will already be clear that the single word can designate a single element of a correlation or two, or more, and even elements in more than one correlation. This depends, as we have mentioned, on the type of language, whether it is isolating, agglutinative, inflected, etc. For example, the English word "men" designates only a correlatum, but the Latin word "hominum" designates both a correlator (with the inflection of the genitive) and its second correlatum, corresponding to the English expression "of men". Certain Italian words, such as "rubameglielo", designate an entire correlational net, corresponding to the English expression "steal it from him for me".

Following the decision to classify the correlational functions of the various contents of thought, by assuming as points of reference the possible correlators, we have prepared a table for each language, called "tabellone", in which the correlators have been listed, leaving the places of the respective correlata empty. Each designatum will have, then, either to be identified with one of these correlators, or constitute one of their correlata. In this connection, the understandable rule obtains that a designatum can never occupy more than one place, more than one box in the same correlation, that is, it is either the correlator, or the first or the second correlatum.

As we have already seen, the mental categories of relationship used by a cultivated person are of the order of several hundreds (see the table which we use for the English language at the Centro di Cibernetica (table 12.2)).

Given the fact that certain designata have many correlational possibilities, as they can act as correlata in almost all correlations, we have examined the possibility of resorting to a negative classification, that is, indicating the correlational impossibilities of the designata. But, at the end, we have seen that, taken on the whole, the two solutions balance each other out.

Having characterized each correlatum by means of each correlator assumed as individual allows us to establish whether a certain designatum can be accepted in that correlation or not, whether with reference to the correlator alone, or to the designatum which occupies the complementary box. For example, the correlation in which the correlator is "between" cannot have as its second correlatum a singular, so that an expression such as "between John" cannot give rise to the correlation in table 12.3.

In order to give it its second correlatum, it is necessary to wait until "John" is accompanied, for example, by a "Mary", when the correlation "John and Mary", classified by number, will come out to be precisely a plural (see table 12.4).

(Even if the rule excludes the insertion of a singular as a second correlatum it immediately shows some exceptions, at least in the Italian language, in which it is possible to say, for example, "fra (una) settimana" (after a week) (see table 12.5) in that "settimana" (week) is assumed as an interval of time, hence with two extremes, the beginning and the end of the week, thus obliging us to prepare a classification in order to isolate this kind of singular from the others. The reference to the complementary correlatum is usually required, at least in many languages, when the correlator does not act as an isolated word, but rather as a suffix, or a prefix, etc., of the word which designates the first or the second correlatum of the particular correlation. Thus we find it for example in the correlations of subject and development, of substance and

Table 12.2

001 shaft	002 aboard	003 about	004 above	005 across	006 afore
007 after	008 against	009 along	010 alongside	011 although	012 amid(st)
013 among(st)	014 and	015 around	016 as	017 aslant	018 astride
019 at	020 athwart	021 barring	022 because	023 before	024 behind
025 below	026 beneath	027 beside	028 besides	029 between	030 beyond
031 but	032 by	033 come	034 down	035 during	036 ere
037 except	038 for	039 from	040 if	041 in	042 inside
043 into	044 lest	045 like	046 mid(st)	047 near	048 next
049 nor	050 notwithstanding	051 of	052 off	053 on	054 once
055 onto	056 opposite	057 or	058 outside	059 over	060 past
061 per	062 qua	063 round	064 save	065 since	066 than
067 that	068 though	069 through	070 throughout	071 till	072 to
073 toward(s)	074 under	075 underneath	076 unless	077 unlike	078 until
079 unto	080 up	081 upon	082 via	083 when	084 whence

Table 12.2
(continued)

085 where		086 whereas		087 whether		088 while		089 whither		090 with	
091 within		092 without		093 yet		094 punctuation marks			095		096
								subject	devel-opment	devel-opment	object
097		098		099		100			101		102
article	. . .	adjec-tive	. . .	. . .	adjec-tive	adverb	. . .	. . .	adverb	predi-cate	. . .
103		104		105		106			107		108
. . .	predi-cate	auxil-iary verb	past part.	auxil-iary verb	pres. part	auxil-iary verb	supine	auxil-iary verb	infin-itive	'to'	supine
109		110		111		112			113		114
'there' 'here'	verb	verb	gerun-dive	gerun-dive	verb	verb	parti-ciple	parti-ciple	verb	. . .	genitive
115		116		117		118			119		120
. . .	dative	vocative	. . .	. . .	voca-tive	com-mon noun	com-mon noun	com-mon noun	proper noun	proper noun	com-mon noun
121		122		123		124			125		126
proper noun	proper noun	X	X	verb	129 infin. acc.	verb	129 sup. acc.	verb	129 finite mood	verb	129 ger. nom.
127		128		129		130			131		132
129 ger. nom.	verb	verb	129 part. nom.	129 part. nom.	verb	129 conj.	129 cond.	129 cond.	129 conj.	. . .	129 finite mood
133 (maintain)		134 (take up)		135 (change of level)		136			137		138
139		140		141		142			143		144
145		146		147		148			149		150
151		152		153		154			155		156
157		158		159		160			161		162

Table 12.3

between	
	John

Table 12.4

between	

and	
John	Mary

Table 12.5

fra	
	settimana

accident, etc., which in many languages are made to agree in number, person, and sometimes even in gender. When there is no element in the form of the words which can be made to agree and which agrees, languages, in order to guarantee the univocality of reciprocal belonging of the two complementary correlata, resort, for the most part, to a quite rigid rule of position for the words which designate them. Another type of agreement between the correlata of the same correlation is that required by the correlators of the group of the so-called conjunctions, such as "and", "or", "but", "although", "though", etc. Here the classification must take into account, for example, whether the correlata be constituted by correlations of subject and development or not, by open or closed correlations, by principal or dependent propositions, etc. Usually, parity of the correlata from these points of view is required; and, among other things, this allows the designation of that which perhaps has been expressed elliptically to be completed. For example, finding "He dresses very elegantly, although poor", since the correlator "although" requires two correlata both with a subject and development, it is understood that "poor" stands for "he is poor". However, this example invites us to reflect further, in that it shows that in these cases we have two alternatives: that which we have adopted, which is to complete the designation, and that of refusing the designatum. Humans have many means to help resolve these dilemmas, suggested both by the way we imagine things, and our culture, to such an extent that perhaps we hardly notice the two possibilities; but the machine possesses none of this, as long as it has not been endowed with it, and so endowing it requires that we first have clarified, that is, individualized, analyzed and described the mechanism which guides the operation of our mind so surely.

I think that the classifications destined to replace the actual operations with which we constitute the contents of thought are now clear; for example, that a thing is liquid or solid, object or activity, that it has only matter or also form, what its dimensions are, etc. Some of these classifications have been collected and presented some years ago (Ceccato [1961]). Table 12.6, for example, is a list of them grouped from the criterion of operatability. These had been prepared for the sentences easily foreseeable on the basis of a modest dictionary, more or less that of basic English. Later we saw that they were not at all sufficient to take the place of the indications which were not supplied by the linguistic expression; but also that their number certainly does not grow in relationship to the size of the dictionary, but, rather, tends to decrease. We also understand how the notional sphere can be of help in guiding the machine to overcome the difficulties of reference or grouping of various thought contents. For example, in the sentence "I have paid for the book with the old prints" it would be very difficult to find a relationship between the notion of paying, book and prints, which would allow us to decide whether the prints were the price paid, that is, the object of exchange, or whether they represent a peculiarity of the book which was paid for; but this is a difficulty for man, and hence for the machine as well. But if the sentence had been "I have paid for the book with the Swiss francs", we would have understood, knowing the relationship between the paying and the francs, that the francs were the price paid, and inserting this relationship in the notional sphere we would have made the machine interpret the sentence correctly. The same would be true if the sentence had been "I paid for the book with the pages missing".

Table 12.7 is a list of relationships which have been taken into account and which constitute the most common cultural heritage. The notions will be found then connected as the example shows. To complete this brief survey of the procedures adopted for machine translation, we add in tables 12.8 through 12.11 some examples of constellations, centered on the

Table 12.6

001 Living beings
002 Non-living beings
003 Animate beings
004 Inanimate beings
005 Human beings
006 Animals
007 Vegetables
008 Minerals
009 Parts of 003
010 Parts of 005
011 Parts of 006
012 Parts of 007
013 Collectives of 002, 004
014 Collectives of 005
015 Collectives of 006
016 Collectives of 007
017 005 + activity (or profession)
018 005 + geographic appurtenance
019 005 + political appurtenance
020 005 + family relationship
021 005 + social relationship
022 Political communities
023 Aquatic animals
024 Flying animals
025 Terrestrial animals
026 Creeping animals
027 Carnivora
028 Herbivora
029 Dangerous animals
030 Peaceful animals
031 Wild animals
032 Domesticable animals
033 Domestic animals
034 Burrowing animals
035 Animals of prey
036 Animals for slaughter
037 Fruit trees
038 Opaque things
039 Transparent things
040 Liquids
041 Aeriform
042 Solids
043 Fluids
044 Powders
045 Fluids and powders
046 Transparent aeriform things
047 Non-transparent aeriform things
048 Transparent solids
049 Celestial bodies
050 Atmospheric agents
051 Atmospheric phenomena
052 Weather conditions
053 Cardinal points
054 Geographic extensions
055 Geographic extensions, land
056 Geographic extensions, water
057 Geographic extensions characterized by their form
058 Geographic extensions characterized by the nature of the ground and of the vegetation
059 Geographic extensions characterized by their relation with extensions of water
060 Foods
061 Solid foods
062 Liquid foods and beverages
063 Foods in powder form
064 Fruit
065 Vegetable foods
066 Natural objects
067 Artefacts
068 Settlements, inhabited places
069 Buildings and constructions
070 Parts of 069
071 Interior parts of 069
072 Exterior parts of 069
073 Objects of interior decoration
074 Furniture
075 Textiles
076 Clothing and things that can be worn
077 Clothing made of textiles
078 Personal effects
079 Parts of 074
080 Parts of 076
081 Instruments
082 Means of transport
083 Means of aquatic transport
084 Means of aerial transport
085 Means of terrestrial transport
086 Domestic utensils
087 Containers
088 Musical instruments
089 Mobile musical instruments
090 Fixed musical instruments
091 Toys
092 Activities and professions
093 Languages
094 Measuring instruments
095 Measures
096 Linear measures
097 Square measures
098 Cubic measures

Table 12.6
(continued)

099	Weights	121	Complementary things for opening or closing something
100	Measures of time	122	Things that can be hung up
101	Names of weekdays	123	Instruments for hanging up
102	Names of months	124	Products of art
103	Indications of time based on astronomic facts	125	Solid things
104	Objects bought with view to something else	126	Hollow things
105	Economic objects	127	Things that can be held in the hand
106	Semantic objects	128	Fixed things that can be held in the hand
107	Events	129	Mobile things that can be held in the hand
108	Public places	130	Things that can be transported
109	Places open to the public	131	Things that can be transported by lifting
110	Places of economic activities	132	Things that can be transported by pushing or pulling
111	Public services	133	Pointed things
112	Activities + their localization	134	Cutting things
113	Things that are covered	135	Things used for indicating
114	Things that can be opened	136	Numbers
115	Covers	137	Things presenting themselves in pairs
116	Things open or closed by adding or subtracting something	138	Things presenting themselves in rows
117	Things open or closed owing to their position with respect to something else	139	Things which owing to their inside play a part in certain activities
118	Things open or closed owing to the position of a part of them	140	Things which owing to their outside play a part in certain activities
119	Sliding or rotating things that open or close something	141	Things which owing to their direction play a part in certain activities
120	Detachable things that open or close something		

Table 12.7

001	element	collection
002	member	class
003	species	genus
004	part	whole
005	component	compound
006	constitutive characteristic	thing characterized
007	subsequent characteristic	thing characterized
008	thing produced	thing which produces it
009	thing produced	place of its production
010	thing contained	container
011	thing supported	support
012	thing pulled	thing which pulls
013	thing directed or guided	thing which directs or guides
014	prominent thing	provenience
015	preceding thing	thing which follows
016	thing covered or closed	thing which covers or closes
017	decorated thing	decoration
018	thing pushed	thing which pushes
019	principal thing	accessory

developments (i.e., the verbs), used principally for translation from Russian into English, Italian and German; these have been taken from a list of 150 (Ceccato [1963]).

[Additional similar tables omitted—Eds.]

The Fourth Approach to MT

We have already spoken of a fourth approach to MT; and the reader will have foreseen what it is. It is a matter of no longer making the machine execute only correlational operations, replacing the constitutive operation of the various thought contents with many classifications made by us and continued by the machine which assumes them as data. This would be the same as endowing the machine with the categorial and representative capacities proper to men, and therefore with the capacity of "seeing" the relationships between things and of reasoning about them without any need of knowing and applying the body of rules which have been obtained from the description of our activities of thought and language, through logic, grammar, rhetoric, etc. It is in the guise

Table 12.8

Russian verb			Operational analysis of development		
BRAT°J			Establishing of relation of appurtenance		
CONSTELLATION					
No.	Relation	No.	Correlation	No.	Contents of the complement
1	The active thing. The dominating term of the relation	1	Subject		
2	The determined term of the relation	1	Object		
		2	Genitive	1	Things measurable quantitatively
3	The point of application of a connection	1	ZA+Acc.	1	Required presence of complement 2.1 and notional sphere relation 'part-whole' with complement 2.1
4	The previous term of relation interrupted	1	U+Gen.		
		2	QT+Gen.		
5	The other term of an exchange	1	ZA+Acc.	1	Notional sphere relation 'part-whole' with complement 2.1 excluded
6	The holding instrument	1	V+Acc.	1	Things used for holding (hand, etc.)
7	The profession	1	V+Acc.	1	Person characterized by profession in plural (the accusative is equal to the nominative)
8	The direction	1	Correlations of directions		
9	Specification of relation	1	S+Instr.	1	Reflexive pronoun (S SOBOJ)

of rulemakers that we find ourselves being classifiers, in botany or in grammar, but not in the guise of simple thinking and speaking beings.

This operation would still lack the fundamental contribution made by memory of the human type and this would weigh heavily on the results; but in part it would be compensated for by this wealth of actual operating which would make the designated things present through semantic connections. We are still however far from thinking of such a machine for MT. At present projects in this direction have a justification only if they are limited to the construction, not of a machine from which practical results too are expected, but of a model for the sake of its own theoretical value, and in which the number of designated things for the time being is very small. With this intent and within these limits [...] we are now actually constructing a machine which observes and describes the events of its surroundings through the actual operations of perception, representation, categorization, etc. But its universe has been restricted to about a hundred designated things (whose combinations are however already quite numerous). In its general lines, the plans of this machine are ready, and its optical device has already been made. Once the entire construction has been completed, it will be possible however to use the machine for an experiment in translation as well, within the range naturally of its very modest dictionary.

An Illustration of the Third Approach

For the moment, then, we are working in the direction we have seen in the third approach to MT: by classifying and making the machine execute only correlational operations, and by integrating the classifications with a notional sphere and some constellations. The best way of showing what the possibilities and the limits of this procedure are is perhaps still that of showing step by step how the machine operates when confronted with a particular linguistic expression; and, as an example, we have used the first sentence of a story for children entitled *The Little Train*. The text to be translated is "Engineer Small

Table 12.9

Russian input verb			Output language	
BRAT°J			English	
Determination of output			Output constellation	
No.	Conditions	Output verb	Comp. No.	Correlation
1		to take	1	Subject
			2.1	Object
			2.2	SOME+Object
			3	BY
			4	FROM
			5	FOR
			6	IN
			7	AS (Replace the plural by singular)
			8	Correlations of direction. See correlational control matrices of the corresponding correlators
			9	WITH+reflexive pronoun agreed in person and number with the subject

Table 12.10

Russian input verb			Output language	
BRAT°J			Italian	
Determination of output			Output constellation	
No.	Conditions	Output verb	Comp. No.	Correlation
1		prendere	1	Subject
			2.1	Object
			2.2	DI
			3	PER
			4	DA
			5	PER
			6	IN
			7	COME (Replace the plural of the input by the singular)
			8	Correlations of direction. See correlational control matrices of the corresponding correlators
			9	CON (Agreement in person and number with the subject)

Table 12.11

Russian input verb			Output language	
BRAT°J			German	
Determination of output			Output constellation	
No.	Conditions	Output verb	Comp. No.	Correlation
1	Absence of compliments 8 and 9	nehmen	1	Subject
			2.1	Object
			2.2	1. ETWAS+Object if singular 2. EINIGE+Object if plural
			3	AN+Dat.
			4	VON+Dat.
			5	FUR+Acc.
			6	IN+Acc.
			7	ALS (Agreement in number with object)
2	Presence of complements 8 or 9 or both	minahmen	1	Subject
			2.1	Object
			2.2	ETWAS+Object if singular EINIGE+Object if plural
			8	Correlations of direction. See correlation control matrices of the corresponding correlators
			9	MIT+Dst. (Agreement in person and number with the subject)

has a little train''. The dictionary is composed of the following words and classifications:

```
1. ENGINEER
   1.1 noun
       gender: masculine and feminine
       number: singular
       common (noun)
       person
       profession
       1.2.1 mood: imperative
             tense: present
             person: II
             number: singular or plural
       1.2.2 mood: indicative
             tense: present
             1.2.2.1 person: I and II
                     number: singular
             1.2.2.2 person: I, II and III
                     number: plural
       1.2.3 mood: supine
             tense: present
```

(*Italian outputs:* for 1.1 'ingegnere', 'fuochista', 'macchinista', 'soldato del Genio'; for 1.2 'costruire', 'organizzare', 'sovrintendere'—as first correlatum of a development-object correlation—'operare come ingegnere', 'operare come fuochista', etc.—in other cases.)

```
2. SMALL
   2.1 adjective
       2.1.1 size
```

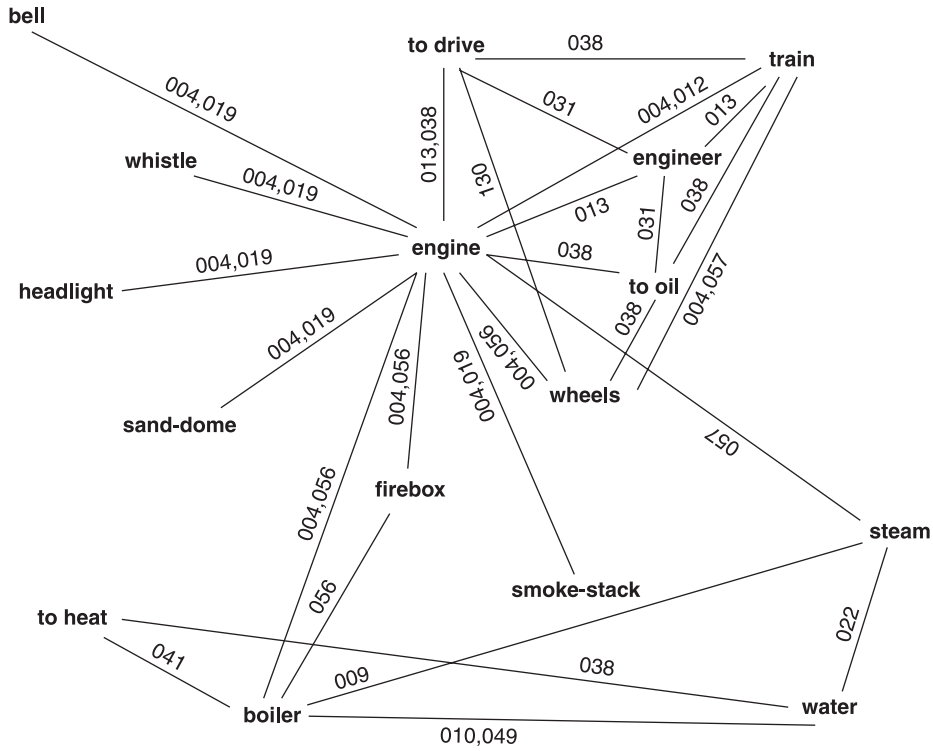


Figure 12.1
From the first ten sentences of *The Little Train*, by L. Lenski.

```
            2.1.1.1 literal sense
            2.1.1.2 metaphorical sense
        2.1.2 intensity
    2.2 adverb
        2.2.1 size
        2.2.2 intensity
    2.3 noun
        gender: neuter
        number: singular
        size
```

(*Italian outputs:* for 2.1.1.1 ‘piccolo’, ‘sottile’, ‘esigu’; for 2.1.1 ‘debole’, ‘legger’, ‘misero’; for 2.1.2 ‘debole’, ‘leggero’; for 2.2.1 ‘poco’; for 2.2.2 ‘basso’ (a bassa voce); for 2.3 ‘piccola cosa’, ‘parte piccola’, ‘parte sottile’’.)

```
3. HAS
        verb
              mood: indicative
              tense: present
              person: III
              number: singular
                        3.1 auxiliary of tense
                        3.2 auxiliary of mood
                        3.3 principal
```

(*Italian outputs:* for 3.1 ‘essere’ or ‘avere’ according to the rules of transitivity, intransitivity, impersonal, reflexive forms, etc.; for 3.2 ‘dovere’, ‘avere da’; for 3.3 ‘avere’, ‘ricevere’, ‘fare’, etc.)

```
4. A
          article
              indefinite
              number: singular
```

(*Italian output:* ‘un’.)

```
5. LITTLE
    5.1 adjective
        5.1.1 size
              5.1.1.1 literal sense
              5.1.1.2 metaphorical sense
        5.1.2 quantity
    5.2 adverb
        5.2.2 quantity
    5.3 noun
        gender: neuter
        number: singular
        5.3.1 size
        5.3.2 quantity
```

(*Italian outputs:* for 5.1. 1.1 'piccolo', 'piccino', 'giovane'; for 5.1.1.2 'meschino', 'gretto', 'futile'; for 5.1.2 'poco'; for 5.2.2 'poco'; for 5.3.1 'cosa piccola', 'cosa piccina'; for 5.3.2 'poco'.)

```
6. TRAIN
   6.1 noun
       gender: neuter
       number: singular
       common (noun)
       thing
               6.1.1 observatum means of
                     transportation
               6.1.2 mental category
                     indicating an ordered
                     succession
   6.2 verb
               6.2.1 mood: imperative
                     tense: present person:
                     II number: singular or
                     plural
               6.2.2 mood: indicative
                     tense: present
                     6.2.2.1 person: I and II
                             number: singular
                     6.2.2.2 person: I, II
                             and III number:
                             plural
```

(*Italian outputs:* for 6.1.1 'treno', 'slitta'; for 6.1.2 'coda', 'fila', 'seguito', 'successione'; for 6.2 'allevare', 'allenare'—as first correlatum of a development-object correlation—'allenarsi', 'esercitarsi', 'viaggiare in treno'—in other cases.)

The words of the text are numbered progressively from right to left, as shown in Table 12.12. The card corresponding to the first word is extracted from the dictionary. Only some of the classifications and the relative numbers from the "Tabellone" are saved and transcribed on a new card. The others are discarded if they are incompatible with the position of the word in the text. In our example, the new card will contain only the classifications of "engineer" as a noun and as a verb in the imperative mood, because of the rule in English which says that the personal forms of the verb cannot appear at the beginning of the sentence. Besides, only some of the numbers of the "Tabellone" assigned to the two classifications which have been preserved will be transcribed: to be precise, for the classification of noun, all first correlata, and of the second correlata only those of the correlations in which an inversion in the order of correlata is possible. For "Engineer", for example, the possibilities

Table 12.12

Engineer	Small	has	a	little	train.
1	2	3	4	5	6 (7)

of second correlatum of punctuation marks, of second correlatum of correlations designated with three words, etc., are excluded. On the new card a new classification not contained in the dictionary is added, that is, the classification of proper noun—a classification which is assigned to all words which in the text are written with the first letter capital, and hence also to the first word of the text and to all the words preceded by a period. The new card of "Engineer" will appear thus:

```
1. ENGINEER
   1.1 noun
       1.1.1 gender: masculine and feminine
             number: singular common (noun)
             person profession
       1.1.2 proper (noun)
   1.2 verb
       mood: imperative tense: present
       person: II number: singular or plural
```

As for the second word of the text, "Small", all the classifications contained in the corresponding card of the dictionary are discarded, as it is a matter of a capitalized word in the middle of a text, and these are substituted by the classification of proper noun, with the numbers of the "Tabellone" which belong to it, except for those which are incompatible with the classification of second word of the text (second correlata of punctuation marks, for example):

```
2. SMALL
       2.4 noun
           proper
```

Then, the correlational possibilities, expressed in numbers from the "Tabellone", are combined according to two general rules: combinations can occur only between numbers whose three first figures (those indicating the correlation) are the same, and whose figures farthest to the right (those indicating the places occupied in the correlation) are complementary. For example, possible combinations are: 001/1+001/2; 001/1+001/3; 001/2+001/3; etc. Combinations of the type: 001/1+001/l, etc., are impossible. Among the combinations possible according to the two general rules, others are discarded in accordance with particular rules for each correlation. These

rules regard: (a) the reciprocal order of the words which designate its correlata in the text; (b) the interval, that is, the possibility that between the words which designate the correlata in a certain correlation there are other words, and what words these might be; (c) the possible agreement (gender, number, person, etc., sequence of tenses, etc.) or other criteria of compatibility between correlata. If, for example, the combination 014/1+014/2 has been executed according to the two general rules, in succession it will be discarded according to the rules of the correlation 014 (correlator ''and'') which requires, in order to be accepted, the order 014/1, 014/3, 014/2, that is, first correlatum, correlator, second correlatum. In our example, the combinations executed according to the general rules of combination and accepted according to the particular rules of each correlation are:

1.1.1 (ENGINEER common noun)
+2.4 (SMALL proper noun)

1.1.2 (ENGINEER proper noun)
+2.4 (SMALL proper noun)

1.2 (ENGINEER imperative verb)
+2.4 (SMALL proper noun)

(The last of these combinations gives rise to the correlation of development object.)

The results obtained are reclassified, this time too according to a group of general rules (for example, certain numbers from the ''Tabellone'' are not assigned to an expression which contains the first word of the text) and particular rules for each correlation. The latter type of rules can be divided into the following four groups: (1) to the correlation are assigned all the combinatory possibilities of its first correlatum; (2) to the correlation are assigned all the combinatory possibilities of its second correlatum; (3) all the possibilities which are common to the first and second correlatum belong to the whole correlation; (4) the correlation has new classifications in regard to those of its first or second correlatum. To the first type belongs the correlation development-object, to the second article-noun, to the third the correlations which have as a correlator categories such as those designated with ''and'', ''or'', etc., that is, the correlations which require parity between the two correlata in order to be executed (a parity which can be different from time to time according to the criteria adopted). To the fourth type belong instead correlations of the subject-development type, which acquire, once executed, different properties from those of their correlata taken by themselves. To these four groups of rules are added their combinations. There will be, that is, some correlations which will be reclassified, according to certain criteria, with the classifications of their first correlatum, and according still other criteria with new classifications. For example, the correlation auxiliary for the formation of tenses—principal verb, of the first correlatum will keep the classifications of number, person, etc., of the second correlatum the classification of principal verb with all the semantic classifications which belong to that particular development, and furthermore, will contain some completely new classifications, such as that of tense. To these rules other restrictive ones are added, when the execution of a combination excludes others of the same sort, or of a different sort. An article-noun correlation will have thus all the classifications of its second correlatum, that is, of subject, object, substantive, for example, but it will no longer retain that of second correlatum of an article-noun correlation, which its second correlatum taken alone had. The reclassifications of the results obtained in the first combinatory cycle are:

for ENGINEER common noun+SMALL proper noun

all the classifications of the second correlatum plus those of the first which regard the fact that we are dealing with a person and a profession;

for ENGINEER proper noun+SMALL proper noun

two reclassifications: one with all the classifications of the first correlatum and one with all the classifications of the second correlatum;

for ENGINEER imperative verb+SMALL proper noun

(correlation development-object) all the classifications of the first correlatum, except the possibility of being again the first correlatum of a development object correlation.

When the reclassification of the results obtained has been performed, these are combined with the possibilities assigned on the dictionary card of the third word, ''has''. (This time there are no reductions in respect to the content of the dictionary card, since these reductions exist only for the first, the second, the next to last and the last word of the text.)

Of ''has'', only the classification 3.3 (principal verb) combines with the preceding constructions, since the other two classifications (3.1 and 3.2) will be put in combination only after they have been combined with principal forms. Besides being combined with these constructions, ''has'' is combined also with the possi-

bilities assigned to the two preceding words taken by themselves without, therefore, taking into account the results obtained.

From this second combination we obtain:

(1) a subject-development correlation which has the correlation common noun-proper noun as a first correlatum;

(2) a second subject-development correlation which has the correlation proper noun-proper noun as a first correlatum;

(3) a third correlation subject-development which has "Small" as a first correlatum.

The input of the personal verb ("has" has excluded the combination in which "engineer" appeared as an imperative), and the recombination of "Small"+"has" with "engineer", not only has not given any result, but it has been discarded because of incompatibility with the preceding text. All three of the correlations obtained in the three preceding combinations are reclassified and supplied with the numbers from the "Tabellone" which belong to them as such. The fourth word in the text, "a", is not combined with the preceding constructions and words, because of a rule of order in English which insists that the article must precede the thing to which it refers. The fifth word of the text, "little", is also the next to the last word. Consequently, some of its correlational possibilities are annulled, for example that of being first correlatum of correlations designated with three words in which it is not admissible for the first correlatum to follow the correlator and the second correlatum ("and", "or", etc.). The classification of "little" as an adjective does not give results, and the same is true of classification as an adverb, whereas the classification of "little" as a noun gives rise to an article-noun correlation, which is reclassified with all the possibilities of "little" as a noun, except, as we have already seen, the possibility of being the second correlatum of an article. "A little" is combined with the results of the preceding combinations and with the isolated words which precede this expression in the text. The following result is obtained (see table 12.13), which in its turn is recombined with the preceding constructions and words giving rise to table 12.14, which will be recombined with the first word, but without giving results (see tables 12.15, 12.16).

All the other combinations executed in these two last cycles have been discarded because of rules regarding the interval between the words corresponding to the correlata. "Train", as the last word in the text,

Table 12.13

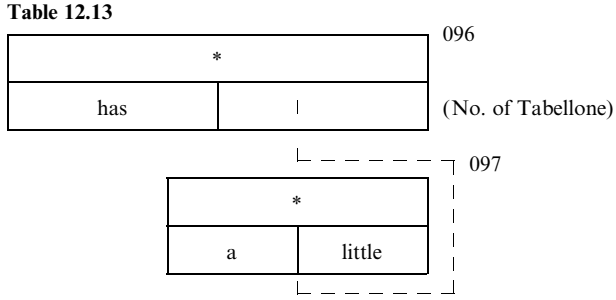


Table 12.14

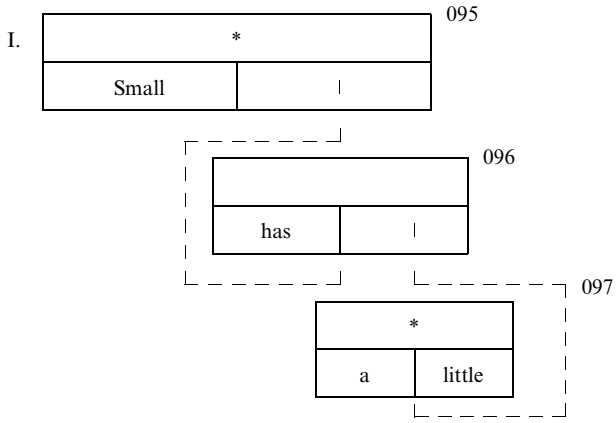


Table 12.15

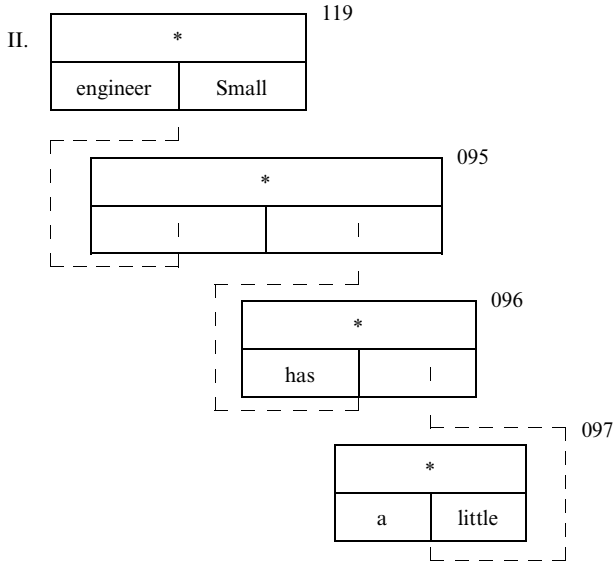


Table 12.16

III.

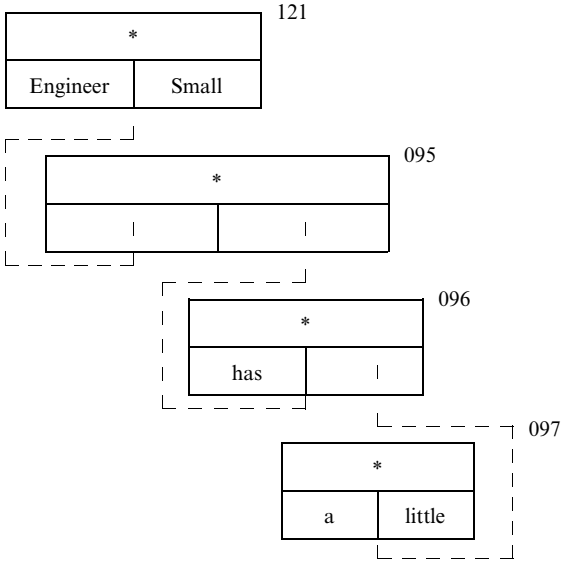
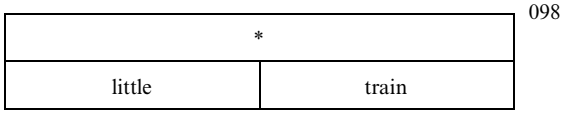


Table 12.17

IV.



retains only some of the correlational possibilities contained in its dictionary card; for example, all its classifications as first correlatum of correlations which do not allow inversions in the order of the correlata are discarded. In addition, only its classification as a noun gives results in the course of the combination, and to be precise, we find table 12.17. The other combinations executed are discarded in accordance with the interval rules.

The result obtained is reclassified with all the combinatory possibilities of its second correlatum, that is, with all the classifications which belong to the noun.

From the succeeding combinatory cycle we obtain table 12.18. Next, table 12.19.

From the cycle in tables 12.20–12.22, the correlational net number VII is combined with the first word of the text, but results are not obtained.

The two final nets, that is, those which contain all the words, number VII and number IX, are compared with a view to the elimination of one of the two. The criteria with which this reduction is performed concern for the most part the examination of the notional sphere, that is, the notional relationships which subsist between their contents. In our example, the relationship is found only for the net number VII to be

Table 12.18

V.

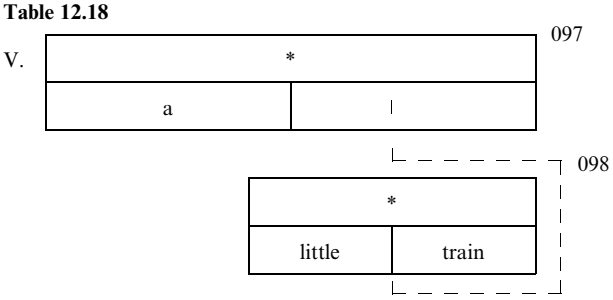


Table 12.19

VI.

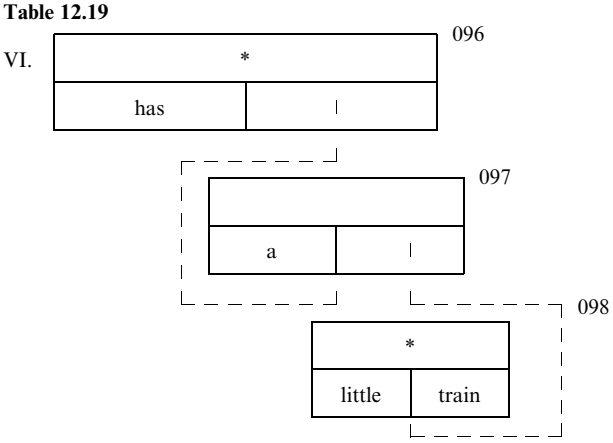


Table 12.20

VII.

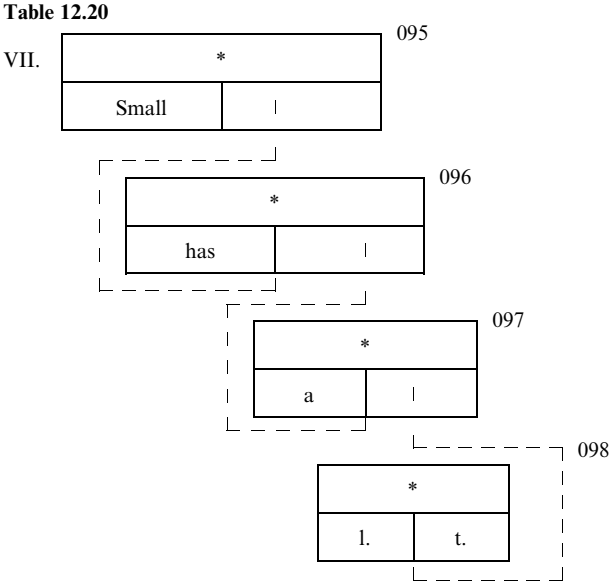


Table 12.21

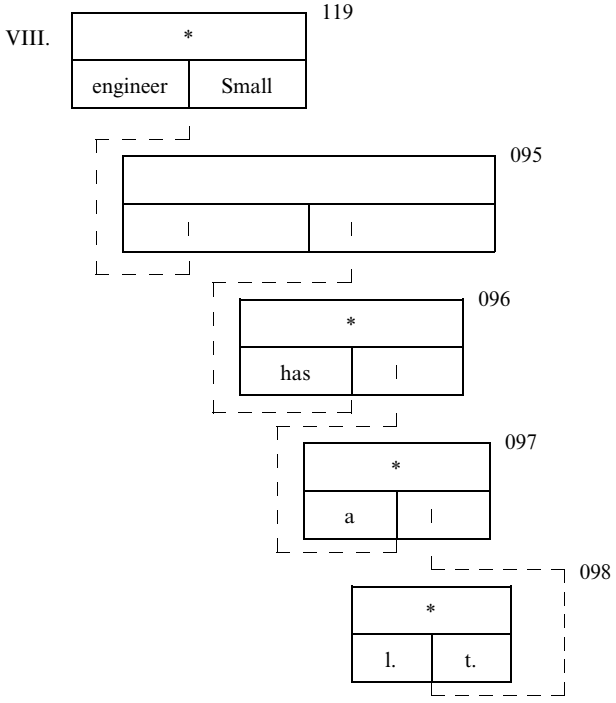
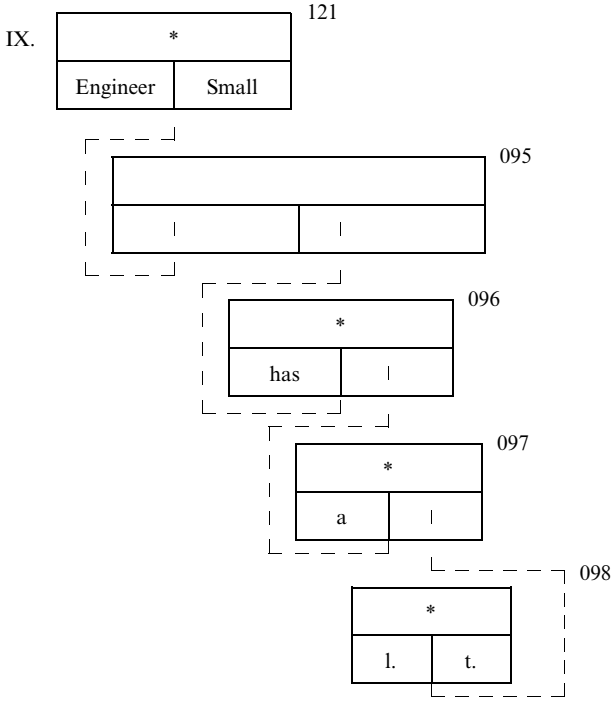


Table 12.22



precise "to direct, guide, or attend to something" (see table 12.20) between "engineer" (with the Italian output "fuochista", "macchinista") and "train" (with the Italian output "treno"). For net number IX no relationship is found, in that "engineer" is here assumed as a proper noun and the Italian output of "train" remains completely undefined. For this reason and for a rule of preference which leads us to choose in the case of the first word of a text its normal accepted dictionary meaning, only net VIII is transcribed for the succeeding operative cycles, that is, for its possible transformation into the corresponding Italian net, and the actual output.

The correlational net is transformed in our example with the addition of a new correlation, that of article-noun (or noun phrase), which will have an article as a first correlatum and the correlation common noun-proper noun as a second correlatum. In fact, in Italian, the name of a profession followed by the name of the person who practices it must be preceded by the article, as long as they do not appear in the correlational net as a vocative or appellation. Since in this case these act as subject, the article must be introduced. Therefore, the transformed correlational net appears as in table 12.23. Now the penultimate phase of the procedure begins: looking up of Italian words in the output dictionary and the choice between them in case of plurality of meanings. As to the choice of the article, definite or indefinite, since in English the equivalent form of "un ingegnere Small" (in expressions as "I met an Engineer Small but I do not know whether it is the same one.") would be a correlation ("an engineer Small", "some engineer Small"), the definite article "il" ("lo", "la", "i", "gli", "le") is extracted from the Italian dictionary.

The choice of output for "engineer" has already been determined by the notional relationships between the correlata of the output net: "macchinista".

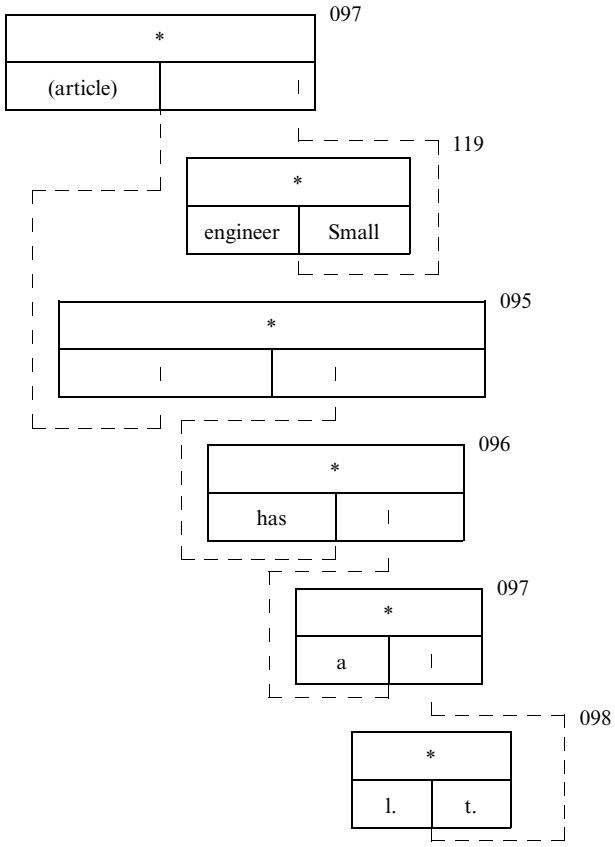
"Small", as it is the proper name of a person, naturally is not translated.

"Has" is translated with "avere", in that for the output "ricevere" a verb form different from the present indicative is required. As to the other outputs "fare", etc., the conditions for the choice (a certain class of objects) do not appear here.

"A" is translated "uno" ("un", "una").

For "little", correlated with a name of an observatum ("train"), that is, with a construction which is constituted also with a figure, the output "piccolo" ("piccoli", "piccola", "piccole"), designating size, is chosen. If "little" were referred to a physical thing (material noun), for example, "butter", that is, to a

Table 12.23



construction to which a figure has not yet been attributed, the choice would have fallen on "poco", designating a quantity.

Also the output of "train" has already been decided, along with that of "engineer" during the choice of the final net.

The last phase of the output regards the inflection of the words extracted from the dictionary. The inflection occurs in successive steps, going left to right and inflecting first the words which in the correlation represent the "dominant" correlatum, that is, first the substantives and then the adjectives, the subjects and then the developments, etc., until all the words in the text have been exhausted. In our example, inflection will occur thus:

stage 1 MACCHINISTA
stage 2 SMALL
stage 3 IL
stage 4 HA
stage 5 TRENO
stage 6 PICCOLO
stage 7 UN

"Il macchinista Small ha un piccolo treno."

References

Brøndal, V., 1943, *Essais de Linguistique Générale* (Copenhagen).

Hjelmsiev, L., 1935, *La Catégorie de Cas* (Copenhagen).

Ceccato, S., 1960, Principles and Classifications of an Operational Grammar for MT, Information Retrieval and Machine Translation, III, 2 (Interscience Publishers, Inc., New York and London), pp. 693–753.

Ceccato, S. (Ed.), 1961, *Linguistic Analysis and Programming for MT* (Feltrinelli Editore, Milano, and Gordon and Breach Publishers, New York), pp. 81–126.

Ceccato, S. (Ed.), 1963, Mechanical Translation: the Correlational Solution, USAF Technical Report, Milano, Ref AF 61(052)-263.

Ceccato, S., 1965a, Automatic Translation of Languages, *Information Storage and Retrieval* (Pergamon Press, Oxford).

Ceccato, S., 1965b, A Model of the Mind, *La Ricerca Scientifica* (CNR, Roma).

IBM Research, 1963, *Final Report for Automatic Translation*, RADC-TDR-63-102 (New York).

International Conference on MT of Languages and Applied Language Analysis, 1961, National Physical Laboratory Symposium No. 13 (Her Majesty's Stationery Office, London).

Tesnière, L., 1959. *Eléments de Syntaxe Structurelle* (Paris).

13
Automatic Translation: Some Theoretical Aspects and the Design of a Translation System

O. S. Kulagina and I. A. Mel'čuk

Eppur' si muove!
—Galileo Galilei

Introduction

The present chapter has two sections. The *first* deals with some considerations concerning the place of Automatic Translation of languages (AT) among problems of wider range: automation of intellectual activities of Man. Three problems are stated on whose solution, in the writers' view, the successful development of AT is largely dependent: the linguistic problem (correlation "text-meaning"), the gnostical problem (correlation "meaning-reality") and the problem of automating scientific research. The *second section* briefly surveys the principal components of an AT system (analysis, semantic dictionary, and synthesis) as it occurs to the present writers, who base their work on the results obtained by Soviet and foreign specialists in this field as well as on their own research.

1 The Place of AT Among Problems of Wider Range

It seems that many (if not all) difficulties in the domain of AT are closely related to, and result from, its following peculiarity (which we expressly formulate here in a paradoxical form): on the one hand, the task of AT as an independent scientific trend is *too broad* and may be quite naturally broken down into a number of simpler tasks which are to be solved autonomously and before tackling AT as a whole; on the other hand, the task of AT, as stated traditionally, is *too narrow* and may be quite naturally included into broader problems which dominate AT in the sense that radical progress in the latter depends to a great degree on the solution of the former.

Translation (not necessarily by machine) is a highly sophisticated operation presupposing: (1) knowledge of the language which is translated as well as the language into which the translation is done, (2) understanding the content of the text(s) translated and (3) knowing how to accumulate translation experience in order to gradually raise the quality of translation.

Let us begin with the first.

1.1 The Linguistic Problem

"Knowledge of a language" probably means the ability to perform two converse operations: transition from text (T) to meaning (M) and, vice versa, from any given meaning to corresponding text(s). By "meaning" we mean here the common content of all texts considered by native speakers to be the equivalent of "what they say" (it should be made clear that this equivalence always holds only with a certain degree of approximation; e.g. ignoring otherwise notable stylistic differences). Thus, the Russian sentences Ваше задание мы выполнили легко ("We fulfilled your task easily"); что Вы нам задали мы выполнили легко ("What you had set us as a task was done by us with ease"); Выполнить Ваше задание нам было легко ("It was easy for us to fulfill your task"); Выполнение Вашего задания оказалось для нас легким ("Fulfilling your task turned out to be easy for us")—have the same meaning. This is exactly how meaning ("*Sinn*") was defined by Gottlob Frege [1892a] more than 70 years ago ("Meaning is what does not change under translation of a text into another language").

For AT needs an algorithmic analogue of this ability to perform the transition from text to its meaning ("$T \rightarrow M$") and vice versa ("$M \rightarrow T$") must be constructed. Calling transition "$T \rightarrow M$" *analysis* and the converse operation *synthesis* we can say that three things are required: a means of recording meaning (a special notation), an algorithm of analysis, and of synthesis. To develop such a notation and to compile the said algorithms are three independent, though closely connected, tasks which, ideally, should have been solved prior to attacking AT as such. Note that all these three tasks emerge in connection with any language information processing (automatic editing and abstracting, information re-

trieval, spoken communication "man-machine", etc.). Small wonder, then, that at present the majority of AT men concentrate on these three separate tasks: automatic analysis, meaning notation, automatic synthesis.

Though, historically, the above tasks have first been faced and strictly formulated within AT, they are, in our opinion, tasks of general linguistics, moreover cardinal problems of any serious theory of language.

Describing language as a communication tool implies, indeed, construction of a functional model imitating human verbal behavior, which, as has been said above, consists in extracting meaning from any given text and generating text(s) conveying any given meaning. Such a conception of the main tasks of linguistics was crystallizing within linguistics itself; but it was only after the appearance of cybernetics and the concept "cybernetical modeling of complicated systems", on the one hand, and substantial widening of the field of applied linguistics, where AT proved to be one of the most illuminating and promising branches, on the other (not to speak about a lot of other important factors, such as the spectacular progress of mathematical logic, improved linguistic methods, etc.), that these tasks could have been conceived as the main tasks of linguistics and given necessary logical formulation. If linguistics had more or less complete solutions to offer here, only some minor (rather technological) problems would have to be solved to make practical AT possible: we would have to record various rules and statements produced by linguists in a way more suitable for modern computers; we would have to take care of economical aspects and ensure sufficient speed of translation procedure; possibly we would have to design special computers for AT purposes, etc. But alas! This is not so. It is just failure of linguistics in solving the said tasks on a strictly linguistic level that constitutes the major hindrance to economical high-quality AT.

Conclusion 1: Any serious progress of AT as a scientific and applied discipline depends mainly on the progress of linguistics in solving these three principal tasks: i.e. to devise a system to record meaning ("a semantic language"), and to compile algorithms of text analysis and synthesis.

It is clear that this progress of linguistics is possible only if linguistics itself is transformed on the basis of new approaches and conceptions, in close connection with mathematics.

Now let us turn to the second problem.

1.2 The Gnostical Problem

Knowledge of a language as stated above ("text → meaning" and "meaning → text(s)") proves in many cases insufficient for different human and/or machine operations on texts. It is well known that a perfect command of the respective languages is not enough for good translation; the translator (or editor) has to perfectly understand what is said in the text under translation, i.e., to have a perfect command of real situations described. Thus, if he deals with a text in a special domain he should be a specialist in this domain. Understanding the "linguistic" meaning of a text does not guarantee the ability to process this text correctly: "linguistic" meaning and "situational" content (the state of affairs) are quite different things not always linked by a unique (one-to-one) correspondence.[1] Let us cite some examples.

A. Different Meanings Correspond to the Same Situation. Consider two groups of Russian utterances:

1. Крупнейший город СССР ("The largest city of the USSR"); Самый большой город СССР (lit. "The most large city of the USSR"); Наиболее крупный город СССР (lit. "The most big city of the USSR"); and

2. Главный город СССР ("The main city of the USSR"); Столица СССР ("The capital of the USSR"); Административный центр СССР ("The administrative center of the USSR").

Within each group linguistically different utterances have the same meaning. To see it, it is enough to know Russian; no information about the extralinguistic world is needed. The utterances in the first and those in the second group have different linguistic meaning: столица ("the capital") and самый крупный город ("the largest city") do not mean the same in Russian; however both meanings refer to the same situations as both groups of utterances describe the same physical reality—Moscow. But to be conscious of it one has to know the real facts, namely, that in the USSR the capital is the largest city. Note that in certain other countries the meanings "the largest city" and "the capital (administrative center)" may well refer to different realities (e.g., in the United States: New York and Washington).

B. The Same Meaning Corresponds to Different Situations. The Russian utterances Для этой цели он использовал книгу ("To this purpose he used the book") and Для этого он воспользовался книгой ("To do this he made use of the book") obviously

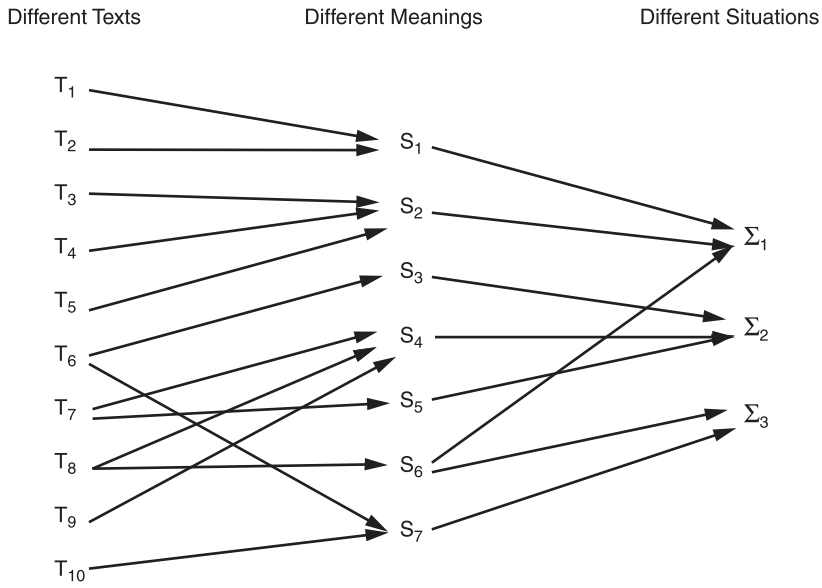


Figure 13.1
Possible correlations between texts, meanings, and corresponding situations.

have the same and quite definite meaning (we emphasize that this meaning is unique: both utterances are not ambiguous!). But this meaning may refer to quite different situations: (1) somebody reads a book to get some information or to divert oneself; (2) somebody puts a book on a ream of sheets, to prevent the wind from scattering them; (3) somebody throws a book at a dog or cat to drive the animal away, etc.

The possible correlations between texts, their meanings and corresponding situations may be schematized as in figure 13.1.

As a rule, there are many texts expressing the same meaning; therefore our main concern is "to try to make out the same thought in its various garments" (Frege [1892a]). The reverse—a text with many meanings—occurs less frequently, and if the text is long enough it is almost excluded. Further, there are usually many meanings referring to the same situation—"many ways to present a situation" (Frege); the reverse is also possible, but not so usual (cf. above).

This distinction—"meaning/situation (reference)"—is not new: it is well at home in logic and logical semantics; linguistics knows it too (Apresyan [1963], p. 106 ff., the opposition "signify-denote", or "meaning-reference" or "intensional-extensional," etc.). Nevertheless, as regards natural languages, neither effective means for autonomous description of (above) separate levels nor rules for transition from one to another have so far been proposed. It is only the text level which has been and remains the object of directed effort to create an autonomous description (various schools of structural linguistics). As for the description of linguistic meaning, to say nothing of extralinguistic situations, here almost nothing has been done. Only recently some researchers have become interested in the problem (e.g., Beltrame [1960], Parker-Rhodes [1961], Ceccato and Zonta [1962], Masterman [1962], Apresyan [1963], Mashinnyi Perevod [1964]). In the majority of linguistic papers the above levels are confused. This is a serious theoretical drawback, but what is more it prevents a number of concrete practical tasks, the task of AT we are interested in for one, from being successfully solved. Unfortunately, even AT specialists did not at first understand the necessity of delimitation and independent study of the named levels. But now it is obvious that if we want our text processing systems, including AT systems, to be *powerful enough*, all of these systems must necessarily contain, besides rules of text → meaning and vice versa transition, rules of meaning → text situation and vice versa transition. Let us give some examples showing how the knowledge of extralinguistic situation can (and must) be utilized in translation.

(1) It is possible that a text is ambiguous, i.e., it has two or more meanings each of which refers to a certain situation, all these situations being different. To choose the meaning intended by the author we cannot rely on our knowledge of the respective language—it

will be of little use; what we need is knowledge of corresponding reality: just this knowledge gives us the ability to decide which of the situations involved is the most probable one. Such is the case with the famous example of Bar-Hillel:

The box is in the pen.

This text has two meanings: "The box is in the small enclosure for children to play in" or "The box is in the writing pen". From the point of view of English both meanings are equally legitimate, but corresponding situations are not. The first is quite natural and highly probable, the second may occur in abnormal conditions only, e.g., in a fairy tale like, say, *Alice in Wonderland*.

The sentence
Slow neutrons and protons

(also stemming from Bar-Hillel) also has two meanings: "Slow neutrons and slow protons" or "Protons and slow neutrons", and both refer to two different situations equally probable from the point of view of physics.

In all similar cases correct translation (which would not be ambiguous as the source text was) is possible only if the following requirement is met: the translating device must be able to get from each meaning to the corresponding situation and then decide which of all situations should be chosen. The first example presupposes many data of the external world (relative sizes of such things as "pen[1]", "pen[2]" and "box", in particular), the second requires not only a deep knowledge of physics, but the knowledge of quite particular physical circumstances connected with the situation implied.

(2) There are other cases where a text has a single meaning which refers uniquely to a certain situation, while in another language, in describing the same situation, its other aspects or features are involved, i.e., other meaning(s) is associated with it.

Consider the following situation: there is a man in a room who hears somebody knock at the door and wants him to come in. In Italian he would say *Avanti!* lit. "Forward!"; in Russian it would be not Вперед! ("Forward!"), which corresponds to the meaning of *Avanti!*, but Войдите! ("Come in!"). Note that *Avanti!* and Войдите! may translate each other only in the situation described. The same goes for many cliché-utterances of this type (greetings, etc.). [...]

In all languages there are many words having very general meaning; to translate such words one has almost always to pass "through" a situation. These words are like French *engager*, *s'engager*, English *get*, etc.

The necessity of using situations is emphasized in Revzin and Rozentsveig [1963], where numerous examples are adduced to show that often the right translation is possible only if the extralinguistic situation is rightly understood.

From what has been said it becomes clear that for AT and, more generally, for any automatic text processing it is necessary to make extensive use of extralinguistic situations and consequently to provide means for a formal description of external world situations and algorithms for transition from meaning to situation(s) and vice versa. Observe that it might be advisable to distinguish not two levels: "meaning" and "physical situation"—but more, introducing intermediate levels which correspond to ever deeper understanding of text and to using an ever broader range of circumstances; in different cases different levels may be used.

The study of correlations between situations (physical reality) and meanings (thoughts about reality) constitutes, in effect, a science dealing with human thinking, with human cognition of the world, with ways the human brain extracts and stores information about this world. Of all real situations only very few (highly special, hardly occurring in everyday practice) are described by exact sciences. However, even in scientific texts, not to speak of fiction or journalism, there are many, in no way special, everyday situations whose description and classification seem to be largely (if not absolutely) ignored so far. It is high time that description of such situations became the object of a special branch of science. In other words, we must proceed to build up a regular encyclopedia of the man-in-the-street's knowledge about the everyday world, or a detailed manual of naive, home-spun "physics" written in an appropriate technical language. It is clear that such a task exceeds the limits of linguistics. So far as the authors know, in AT and related fields only first steps have been made in this direction (cf., especially Mashinnyi Perevod [1964], "Foreword" by A. Žolkovsky).

To sum up: We establish a clear distinction between meaning level and situation level. Just as linguistics undertakes to devise a semantic language and "text $\rightarrow$ meaning" ($T \rightarrow M$) and "meaning $\rightarrow$ text(s)" ($M \rightarrow T$) algorithms, meaning being recorded in the said semantic language, so some special discipline (maybe "gnostics"?) should tackle the task of developing a "situational" language and

"meaning → situation" (MS) and "situation → meaning" ($S \rightarrow M$) algorithms.

Conclusion 2: Any substantial progress of AT is closely dependent on progress in the study of human thinking and cognition, in particular—on the successful solution of such tasks as developing a formal notation for recording external world situations and constructing models of thinking (meaning analysis and synthesis). Since the object of AT—at least, in the foreseeable future—is mainly scientific and technological texts, what has been said refers not to the astronomical variety of all conceivable situations, but to the rather limited (though still very large) range of situations possible in special fields.

We shall now dwell [. . .] on the third problem.

1.3 The Problem of Automating Researchers' Activity

The AT system containing all four algorithms: $T \rightarrow M$, $M \rightarrow T$, $M \rightarrow S$ and $S \rightarrow M$ will necessarily have an enormous volume and a very complicated logical structure. It should have at its disposal a complete set of data about two (or more) languages: lexical, syntactical, stylistic information and the like, data about distribution and functioning of all items in the whole range of possible contexts in the respective languages, rules of correspondence between these items, etc. Besides, it should store an extremely large amount of data on the situations described, in other words—a real encyclopedia. More still, all this information should be organized in such a manner as to exclude contradictions, inconsistencies, incompleteness, etc. The problem is complicated by the fact that all these data change in the course of time, and the data concerning situations may change rather quickly.

It is highly improbable (at least, in our opinion) that an exhaustive and consistent system of this kind could be constructed by a staff of researchers from beginning to end "at one go." At a certain moment such a system becomes so complicated that control, corrections and additions exceed the limits of human abilities. No scientist or group of scientists is able to keep up with the growing system as a whole and do the necessary bookkeeping. We believe that the said difficulty can be overcome only by automation of the procedure for correcting and updating the algorithms. It seems that we should begin with building up an approximate AT system—it may be far from complete, contain errors and inconsistencies, but it must be easy enough to control. This system should be provided with some "maintenance" means: a set of algorithms controlling the present state of the system, ensuring the harmless introduction of new data (which must not contradict the information stored in the system) and—what we believe most important—providing the researcher with any information he may need about the internal state of the system at any given moment.

Such a system can function as follows: a completed variant of the system is set to work. The results are examined by specialists, errors are marked off and classified, and necessary corrections and additions are suggested which are introduced into the AT system by means of the said maintenance devices. These maintenance devices should be self-improving so that they take over an ever greater part of the effort expended on improving the AT system. Not only should the maintenance system introduce into the AT system the man-made corrections and additions, but it should also "learn" to analyze of itself the translation errors and defects marked off by the human editor, to formulate independently hypotheses about the necessary corrections in the rules involved, to check them, also independently, by asking man when necessary, and to incorporate these corrections into the AT system.

Along with such a maintenance system we should devise systems for automatic collection and classification of language data suitable for processing any kind of text, that is, prepared beforehand (e.g., with syntactic relations marked by a linguist) or unprepared. The products of such systems may be used by linguists as well as by maintenance devices of AT systems.

While an AT system is a model imitating the activity of the human translator, the system for collecting language data plus the maintenance system is a model of the activity of the scientist devising the former model. Some preliminary steps have been made in this direction (Harper and Hays [1959], Giuliano [1961], Kulagina [1962, 1963], Progress Report No. 1 [1962], Final Report No. 16 [1963]).

Similarly, the activities of programmers engaged in programming an AT system should be automated as fully as possible. We mean the automation of programming as such, of optimizing the arrangement of the data in the storage device, of coding etc.

As a matter of fact we are speaking here about automating the research activities of Man, i.e., about problems in the same area as automatic pattern recognition, learning, logical deduction and the like—in short, problems of artificial intelligence.

Conclusion 3: The practical solution of the AT problem depends, to a great degree, on our ability to automate the scientific activities of humans (heuristics).

Thus we believe that the future of practical AT depends upon the solution of three types of problems: the linguistic problem (creation, with the help of mathematics, of systems of formal description of natural languages); the "gnostic-encyclopedic" problem (development of formal descriptions of human knowledge about the external world); and the "logical-heuristic", or artificial intelligence, problem (formalization of thinking processes). One should not, however, be too pessimistic and freeze all research in the field of AT till the above problems are solved. On the contrary, we believe that work on and in AT should be developed in its theoretical as well as practical and experimental aspects. Firstly, AT seems to be a meeting point of all the three problems. Therefore it is an area whose development, in our opinion, is essential for the solution of those problems. Theoretical suggestions are checked practically in AT; on the other hand, it is in this field that many new tasks have emerged and new scientific problems have been first formulated. Secondly, AT cannot be reduced to the three problems just mentioned; besides these, AT has to deal with a number of other problems connected with constructing complicated cybernetic systems of special types. It is necessary to know what pecularities are characteristic of AT systems, to accumulate experience in designing this kind of systems, to trace out and elaborate their principal parts. In a word, our way should be a sequence of approximations (let us hope this sequence is a convergent one!). Along with elucidating theoretical questions we should create functioning AT systems and proceed from theoretical studies to experimental tests and back again. In this connection we should like to call attention to the importance of large-scale computer experiments in AT. Here positive results obtained by some American AT groups should be noted.

2 Principal Components of an AT System

We want now to sketch an AT system of a type which seems feasible at the present time. In other words, we mean to propose a minimum program which probably can be realized before the above theoretical problems have been completely solved. It should be made clear that we shall restrict ourselves to "translation *via* meaning" (provisionally, without referring to situation), for we believe that such translation is realizable now, and despite all its drawbacks it can attain a much higher level of quality than hitherto achieved by any other AT.

The proposed AT system consists of three main components: (1) an analysis algorithm, (2) a semantic dictionary, (3) a synthesis algorithm.

2.1 Analysis Algorithm

The analysis algorithm is composed of: (1) lexico-morphological, (2) syntactic, and (3) semantic, subalgorithms.

2.1.1 Lexico-morphological Analysis This algorithm aims at assigning to each occurrence (word form) in the text a special code, called *information* (technical term), which represents data of two types: syntactic information concerning possible distribution (combinatory properties) and potential syntactic functions of the given word form, and semantic information describing its meaning in terms of a set of semantic elements and parameters fixed beforehand. This algorithm operates with a list (dictionary) of morphs of the given language; each morph is assigned "partial information." It also uses rules of morph combination and rules of conjoining partial information elements assigned to morphs into "definitive (complete) information" assigned to the whole word form (string of morphs). Different methods of lexicomorphological analysis are described, e.g., in Zasorina et al. [1958], Oettinger [1960], C.N.R.S. [1961], Dupuis [1961], Mel'čuk [1961], Blois et al. [1963]. This assignment is not unique: a word form may be assigned various information elements standing in relation of (strict) disjunction (homonymy, or homography) and/or of conjunction (compound words); many different word forms may be assigned the same information (e.g., Russ. книгой and книгою ["by the book"]).

2.1.2 Syntactic Analysis Syntactic analysis associates with each sentence of the text all permissible (regardless of semantics) structural descriptions, or syntactic structures (ss) called *syntactic trees* (we have agreed to represent the syntactic structure as a tree in the sense of the theory of graphs). The input to syntactic analysis is the string of information elements assigned to word forms of the sentence analyzed; its output is the tree of the sentence, a diagram representing syntactic connections between word forms and between parts of the sentence (if the latter consists of more than one clause). A string of information elements (i.e., the input sentence) may be assigned several trees, which is a manifestation of syntactic ambiguity. Different strings of information elements

may be assigned the same tree. In the course of syntactic analysis many cases of morphological (word form level) homonymy are resolved.

Two aspects of syntactic analysis should be kept apart: the method of representing the results of analysis (the syntactic structure, when it is found) and the method, or strategy, of obtaining these results, i.e., of finding out the structure (Mel'čuk [1963b]). As regards the former, in AT—and more generally, in linguistics—two main approaches are known: "arrows" and "brackets," or in more technical terms, *dependency tree* and *constituents tree* (the analytical collation of the two methods is given in Hays [1961] and Paducheva [1964]). It seems reasonable to combine both of them. As a rule we use dependency trees ("arrows"), but in many cases where this representation is not sufficient: e.g.,

The tasks and methods of mathematics = (the t. and m.) of m. or *the t. and (m. of m.)*

We feel it is necessary to make use of constituents trees too.

Turning to the methods of determining the syntactic structure we should distinguish two approaches: sequential analysis and filter method.

Sequential analysis, despite the multiformity of its technical realizations [predictive analysis of I. Rhodes and A. Oettinger (Alt and Rhodes [1962], Bossert [1960], Bossert et al. [1960], Oettinger and Sherry [1961], Rhodes [1959]), fulcrum analysis of Garvin [1961], algorithms of Harper and Hays [1960], Corbé and Tabory [1962], Moloshnaya [1960], Mel'čuk [1963b, 1964] etc.], boils down to the following. The algorithm describes the *process* of establishing the syntactic structure. All words are classified according to some predetermined criteria and corresponding class marks are put in the information assigned to each word. For each word class a special instruction (routine) is given which leads to determination of syntactic functions (in the sentence analyzed) of any word belonging to this class. This routine essentially relies on any information about the functions of other words which has been obtained by the time the given word form (i.e., its information) is processed. When the syntactic function of a word form is established the information assigned to it may be modified and thereupon this word form goes with this modified information. It means that results of processing n word forms are essentially used when the $(n+1)$th word form is processed. Under such a method it is possible that, before the analysis is finished, there are some words for which authentic definitive syntactic connections are established, some other words for which only hypothetical connections could be established and whose function will be definitively fixed later, and words which have not been processed at all.

The *filter method*, or its idealized scheme which is modified when programmed on a real computer, does not describe the (dynamic) process of finding out the structure; instead it uses some *static* description of all regular (well formed) syntactic structures of the given language, i.e., certain conditions or requirements imposed on a supposedly well formed structure. Since syntactic structure is a set of connections ("arrows") between text elements, the process of finding it is quite trivial here: the algorithm forms all possible combinations of binary connections between elements—all hypothetical syntactic structures, or HSS's (the string of three elements *a b c* leads to 9 HSS's:

a b c, *a b c*, *a b c*, *a b c*, *a b c*, *a b c*, *a b c*, *a b c*, *a b c*,

which are checked for meeting the conditions of well formed syntactic structure), and after all HSS's have been sorted out, all those (and those only) are left as correct ones which satisfy all conditions. We see that each condition works as a filter rejecting a number of wrong HSS's. When the filter method is adopted the algorithm works stage by stage each of which constitutes the application of a filter to all elements of the input sentence or to this sentence as a whole. The sequence in which words are processed within a stage is irrelevant. Words are processed at a more even pace: it is impossible that all connections of some words should have been established while other words have not yet been processed at all. In practice a number of conditions are taken into account in the process of forming HSS's in order to minimize the sorting of all HSS's (Lecerf [1960, 1961], Slutsker [1963], Iordanskaya [1964]).

It seems that an optimum can be attained if both methods are combined in such a way that sequential analysis considerably reduces the number of HSS's to be filtered, provided this combination does not make the algorithm too complicated.

It should be emphasized that whatever the general method of finding ss, we are sure that it should be elaborated on the basis of static description of the regular ss in the language under consideration, i.e., by way of fixing properties of the regular ss independently of processes conducting to its determination. Such an approach, besides purely theoretical

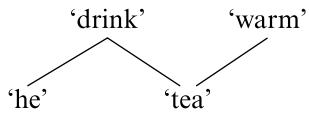
advantages, is convenient in terms of organization of work, for it permits us to isolate two different tasks: the linguistic task of describing ss properties in a given language and the mathematical task of constructing algorithms for finding ss's possessing certain well defined properties. It enables linguists to make full use of their professional knowledge and study properties of linguistic objects without bothering about algorithmic problems alien to them, and mathematicians—to concentrate on devising economical algorithms for finding rigorously defined regular ss's.

One should not forget that an analysis algorithm which would guarantee for any given sentence all right ss's and only those is practically unfeasible. Any algorithm will make mistakes: either admit wrong ss's, or lose some right ones. So we must make our choice in advance: we may want either maximal completeness of the algorithm (all right ss's are found) or maximal adequacy (all wrong ss's are rejected).[2] In the first case we run the risk of having among right ss's some wrong ones, in the second—of rejecting some right ss's. This problem is analogous to that of the automatic sorting of articles in industry where one has to choose between rigid sorting (risking sorting out some good articles) and loose sorting (risking leaving some deficient articles).

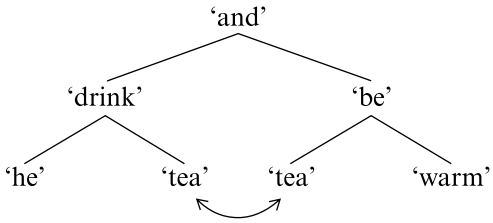
Since it is impossible to attain absolute completeness and absolute adequacy, either together or separately, it seems reasonable, in developing an algorithm, to aim at completeness in order to obtain, for as large as possible a set of sentences, all or at least some right ss's; we are ready to admit among the latter some wrong ones. When our algorithm attains the level of completeness planned in advance we can raise its adequacy by making the description of ss properties more accurate and/or introducing finer filters.

2.1.3 Semantic Analysis[3] Semantic analysis provides for each ss a semantic structure (SEMS). The input of semantic analysis is a syntactic tree and semantic information about its nodes, the output—the SEMS of the sentences, that is, a set of semantic connections between elementary (in terms of the system adopted) semantic items. A syntactic tree may be assigned several SEMS's (semantic ambiguity of the input ss), and several syntactic trees may have the same SEMS. Semantic analysis may lead to the resolution of syntactic homonymy, i.e., to eliminating syntactically possible trees with unacceptable semantic interpretation (cf. the structure ($4'_E$) in figure 13.5.

In semantic analysis, just as in syntactic analysis, we must distinguish, between questions of representing SEMS's and of determining them. As regards representation, we also may use trees. Thus, the method of representing SEMS's used in Žolkovsky et al. [1961] and called predicate notation (предикатная запись) is a linear notation equivalent to a tree diagram. There are, however, some difficulties resulting, among other things, from the existence of modifiers in natural languages. In such a (quite simple) sentence as *He drinks warm tea* the word *tea* is semantically dependent[4] on two predicate words: *drinks* and *warm*, which is prohibited for a tree (the structure



is not a tree[5]). This sentence may be assigned a different and more sophisticated semantic tree-diagram:



which corresponds to the semantically equivalent sentence *He drinks tea and this tea is warm*. The latter diagram marks semantic elements referring to the same object of reality by means of a bold type arrow.

Other ways of representing SEMS's are also possible. As in the case of ss's, we consider as optimum a combination of different ways each of which represents a definite aspect of semantic connections and which, taken together, characterize the SEMS as a whole just as different projections of the same object characterize it on a drawing.

For lack of experience, nothing definite can be said about methods of deducing SEMS from ss and semantic information in its nodes. We may hope, though, that two strategies analogous to sequential analysis and filter method in syntax, will be distinguished here too.

Considering translation "via meaning" (situation disregarded) we assume that SEMS ($\simeq$meaning) obtained by analysis of input text may be the starting point for synthesis of text in the output language, which boils down to identifying SEMS of different languages. We have said in the first section that doing so can result in errors or ambiguities in the translation

output. To ensure reliable and smooth translation, the stage of semantic transformation (SEMS of language L1 → SEMS of L2) realized by means of "descending" (or "ascending"?) to the common "situational" structure is quite indispensable. But this stage being only a desideratum (no practical work has been done here so far) we shall review our translation system without it and assume that text analysis ends with establishing the semantic structure.

Synthesis immediately follows.

Parallel to the three stages of analysis as described above, there are three stages of synthesis in inverse order: semantic synthesis (from SEMS to all possible —from the point of view of the output language— syntactic trees), syntactic synthesis (from a syntactic tree to all possible strings of word information elements giving corresponding output sentences) and lexicomorphological synthesis (from information to the real output word in the conventional graphic form; each string of information elements is transformed into an output sentence).

When passing from a string of words to its ss (analysis) and vice versa (synthesis) the same lexemes (with the exception of auxiliary elements) which occur in the input text or are to constitute the output text (see examples) are kept in the nodes of the syntactic tree. But when passing from a ss to the corresponding SEMS the nodes undergo certain transformations. All input lexemes are replaced by semantic items, which will be considered below. In some cases this replacement is biunique—technical terms and such term-like non-ambiguous words as *red* (color) → 'red'[6], *car* → 'car', *eye* → 'eye'. In other cases various transformations of nodes take place, namely: 1) several nodes are replaced by a single node одержать победу 'gain a victory' → 'победить', 'id'., *inflict a defeat* → 'defeat', железная дорога → 'railway'); 2) a node is replaced by several semantic nodes (мчаться 'rush along' → 'very/great'+'fast'+'move'; мы обозначаем индекс буквой 'we designate the index by a letter' → ['we'] + 'make' + ['letter'] + 'designate' + ['index']); 3) a syntactic node is replaced by a single or several semantic node(s), but this replacement is not unique тяжелая рана 'a grave wound' → ['wound'] +'great degree', тяжелый камень 'a heavy stone' → ['stone']+'very/great'+'weight').

The transition from a SEMS to its ss's under synthesis implies analogous, but inverse transformations: (1) 'help' → [to] *help* or *give help*; (2) ['success']+ 'great degree' → [*a*] *dramatic* [success]; (3) 'staff' → *staff* [*laboratory*], *personnel* [*hospital*], *crew* [*tank* or *ship*], *team* [*football*], *troupe* [*theater*], etc.

All these transformations are carried out with the help of a semantic dictionary which constitutes an integral part of the translation system.

2.2 Semantic Dictionary

The semantic dictionary is based on the following considerations (research along similar lines is described in references mentioned in note 3).

Translation, as we understand it, presupposes the transition "text → meaning" and vice versa. The special language necessary to record the meaning of the input (and output) text (*pour fixer les idées* we shall consider English into Russian translation) may be conceived as a kind of standardization, or simplification of the input and output languages; something like Basic English, resp. Basic Russian. We abstract ourselves completely from the grammar of this language. As regards its vocabulary, it consists of a limited set (a few hundreds of items plus technical terms) of English (resp. Russian) words chosen in such a way as to permit us to express the meaning of any given text. All terminological words are simply included in this "Basic", and other words of the input language must be expressable in Basic by means of non-ambiguous and readily understandable paraphrases, no matter how clumsy and unnatural they may seem (see examples at the end of the final section). Having such Basics one may conceive of translation as of a three-stage operation: from English into Basic English (transition from English text to semantic structure), then from Basic English into Basic Russian and, finally, from Basic Russian into good, idiomatic Russian (transition from semantic structure to Russian text). It seems obvious that due to the limited vocabulary and elementary character of its items, translation from Basic English into Basic Russian would be much simpler than translation between two idiomatic languages. Maybe, it would be still better to eliminate this operation by merging the two Basics into one.

In the future this merger Basic should be amplified, to match a larger range of natural languages, and made more precise so that each of its elements is assigned one and only one strictly fixed meaning. It is this perfected Basic that will be Semantic Language, or AT Interlingua (язык посредник).

This approach enables us to divide up our problems and tackle the difficulties separately. It becomes possible to distribute work between independent research groups with each of them concentrating on a very simple and concrete task: e.g., to choose English words which are necessary and sufficient to paraphrase a number of given texts, etc.

English-Semantic Zone	Semantic-Russian Zone		
	a) Russian of degree zero	b) Russian of degree one	c) Russian of degree two, etc.
x → α yz → ε w → βαγ v → (1) β (2) δ	α → А β → БВ γ → (1) БД (2) ГЖ δ → В	А → (1) К, (2) Л, (3) М БД → Н В → (1) У, (2) ОПР ГЖ → СТФ	. . .

Figure 13.2
Scheme of semantic dictionary. Roman letters denote English words; Cyrillic letters, Russian words; and Greek letters, elements of semantic language.

From what has been said above follows that an English-Russian AT dictionary should be divided into two zones: 1) English-Semantic, 2) Semantic-Russian. The second zone, in its turn, may be divided into several sections: the first ensures the transition from Semantic to highly simplified Russian ("Russian of degree zero"); the following, the transition to more and more idiomatic and rich Russian ("Russian of degree one, two ...", etc.—*ad libitum*). The number of these sections depends only on the degree of elaboration of the whole system. See figure 13.2.

2.3 Synthesis Algorithm

The main phases of synthesis have been enumerated above: semantic, syntactic and lexico-morphological synthesis.

We shall dwell in some detail on syntactic synthesis. About semantic synthesis nothing definite can be said so far (however, see Žolkovsky and Mel'čuk [1965]; lexico-morphological synthesis is described in a number of papers, e.g., Volotskaya [1961a,b], Foust [1960], Walking [1960].

Syntactic synthesis consists in transition from the syntactic tree of the output of this sentence. That is, syntactic synthesis determines the form (grammatical characteristics) of output lexemes and their arrangement (word order). For more details on syntactic synthesis see Mel'čuk [1965].

Syntactic synthesis of a complex sentence is carried out first on its parts—clauses and clause-like stretches (participle, gerund and absolute constructions [...]. Each part is synthesized separately, and only then a special routine makes the assembly according to the structure of these parts and the type of syntactic connection between them.

Syntactic synthesis of a clause (простое предложение) includes three steps. The first step successively isolates, in the tree diagram of the sentence, subtrees corresponding to *Primitive (Initial) Word Groups* (PWG). Each PWG consists of a head and some of its immediate "slaves" (dependents). In synthesis a group is dealt with as a monolith. In the course of rearrangement of the sentence according to word order requirements it is moved (e.g. is inserted within other groups) as a whole, without changing the order of its members. There are four group patterns: verb group, noun group, adjective group, adverb group. For instance, the Russian noun group pattern is as in table 13.1.

In the synthesis algorithm position and grammatical characteristics of each member of the group are fully determined by the three following values: wordclass of the head, wordclass of the "slave" and the type of syntactic connection (SCT) between them.[7] These values are utilized in a three-dimensional matrix called *Group Grammar*. SCT gives section k of the matrix, the wordclass of the head gives column i, and the wordclass of the "slave" gives row j. Case $A_{k,i,j}$ contains the code describing the relative place of the "slave" (e.g., possessive adjectives like мой, 'my' or их 'their' take the fourth place to the left of the corresponding noun—see above) and its grammatical form (possessive adjectives like мой ("my") or ваш ("your"), but not его ("his"), ее ("her"), or их ("their"), should have the same gender, case, and number as the noun head of the group). Note that for members of a word group, the form and position in the text in regard to the head of the group is given independently from other "slaves" of the same head and/or from other word groups.

Table 13.1

1	2	3	4	5	6	7	8	9	10
Conjunction	Negation	Restrictive particles	Preposition	Quantifier (cardinal numeral etc.)	Demonstrative	Possessive adjective	Ordinal numeral	Adjectives	Noun (head of the group)
и	не	только	для	двух некоторых	этих	наших	первых	важных	решений
And	not	only	for	two some	these	our	first	important	solutions

The second step consists in assembling all the PWG's into *Definitive (Terminal) Word Groups* (DWG). All PWG's are tested in turn for their "masters" (the "master" of a word group is that of its head). All PWG's with "masters" other than the sentence predicate (i.e. the top node of the syntactic tree) are included, in a special way, into the group of their "master" to form DWG's. The type of the PWG and the type of its syntactic connection with its "master" uniquely give its position relative to its "master" group. Two cases are possible: (1) A group must be inserted (nested) within its "master" group. Thus, the adjective group связанный с указанной задачей "related to the said task" may be inserted in the "master" group в наш алгоритм "in our algorithm" immediately before the noun to produce a DWG в наш связанный с указанной задачей алгоритм lit. "in our related to the said task algorithm". (2) A group must be put immediately before or after its "master" group. E.g., the noun group ... автоматического перевода "of automatic translation" is put to the right of its "master" group ... в наш алгоритм resulting in the DWG в наш алгоритм автоматического перевода.

If two or more groups must be put on the same side relative to their common "master" group, special rules are introduced to decide their mutual order. Such rules take into account grammatical norms of the output language, the simplest stylistic factors (e.g. the length of dependent groups) and (in order to avoid ambiguities) some considerations of meaning. Thus, assembling the groups 1) с помощью указанных наборов правил ("by means of the said sets of rules"), 2) на заданные классы ... ("into given classes") and 3) всех таких преобразований ("of all such transformations"), with the "master" group первое разбиение ("the first breaking down"), the algorithm must produce the DWG первое разбиение всех таких преобразований на заданные классы с помощью указанных наборов правил (lit. "the first breaking down of all such transformations into given classes by means of the said sets of rules"), and not первое разбиение с помощью указанных наборов правил всех таких преобразований на заданные классы, which is ambiguous; either "breaking down rules of all such transformations by means of the said sets" or "breaking down all such transformations by means of the said sets of rules".

The second step results in an array of definitive word groups. These DWG's are: the finite verb group; the groups of subject and objects ("actants," to use Tesnière's term, Tesnière [1959]); the groups of circumstantial complements ([...] "circonstants," in Tesnière's terms); adverb groups modifing the finite verb; and groups of nominal or infinitive complements of a compound predicate [...].

The purpose of the third step is to arrange all the resulting DWG's to ensure an acceptable Russian word order. In contradistinction to the word order within PWG's and to the order of PWG's within DWG's, the mutual order of DWG's is not uniquely dependent on the properties of two connected DWG's and on the type of syntactic connections between them (at least, in Russian). To define the right order, not only the properties of each pair of the DWG's and the corresponding connections must be taken into consideration, but the whole set of all present DWG's as well: presence or absence of certain DWG's, co-relative properties of many DWG's considered simultaneously, presence and localization of *logical emphasis* ("актуальное членение"), and many other closely interwoven factors. The algorithm we propose for the arrangement of the DWG's does not take care of all these factors simultaneously, but instead operates as a sequence of five individual devices.

The first device makes a hypothetical arrangement of the groups of (finite) verb and "actants", considering only some properties of the verb and the presence

of such and such "actants". As a result something like the skeleton of the output sentence is produced; if there were no other factors (other groups), this arrangement would be all right. The successive devices will gradually "cover the skeleton with flesh", changing the skeleton itself where required by new factors.

The hypothetical skeleton can be divided into different kinds of components according to the degree of their susceptibility to later modifications. Some components are strictly fixed, that is, their mutual order cannot be altered by any factor. Thus, in Russian for a subject group or nominal predicate group expressed by a noun when the copula is zero, only the direct order is possible (in scientific texts): мой друг хороший врач ("my friend is a good physician"). Other components are not fixed, their order in the skeleton may be modified by certain "strong" factors, e.g., logical emphasis; instead of the usual существует такое отражение ("There is such a mapping") we can have, if существует "is, exists" is to be emphasized, такое отражение существует "Such a mapping does exist". Still other components may be affected by numerous "weaker" factors, etc.

To take these differences into consideration, the concept "weight of an arrangement" is introduced. The *weight* of a given arrangement is a number characterizing the degree of fixedness of the word order (i.e., of the order of DWG's) within individual sections of the sentence skeleton, or the degree of fixedness of an element in a given position in the skeleton. These numbers have been chosen quite empirically to represent the relative strength of individual factors according to the intuitive feeling of the researchers. Thus, the sequence "subject group+nominal predicate (the copula being zero)" is rated as the maximum weight, say 5; the sequence "finite verb of existence+subject group" имеется алгоритм ("there is an algorithm") is assigned some lesser weight, say 3, and so forth.

Each factor considered by the successive devices and requiring rearrangement of the groups previously located is given a tag indicating the weight this factor is able to overcome (outweigh) and the weight the sequence resulting from rearrangement must be assigned.

The second device adds to the skeleton all circonstants according to the skeleton type and properties of the present circonstants. Note that the arrangement of the actants and the distribution of the circonstants among them is oriented on some "normal", or neutral, word order; this is the word order a native would choose facing the given set of actants and circonstants if the arrangement were not affected by other factors (if the output sentence were deprived of logical emphasis, all definitive word groups were of equal length, all ambiguities were excluded, etc.). Examples of neutral word order are; мальчик идет в школу ("The boy is going to school"); я читаю книгу ("I am reading a book"); возникают следующие трудности ("The following difficulties arise"); В лаборатории изучаются сложные процессы ("In the laboratory complex processes are studied"), etc.

All circonstants located, too, are assigned a certain arrangement weight.

The third device takes into account all kinds of logical emphases (which are considered to be meaningful elements; the corresponding information must be generated by the analysis of the input text and carried over into the synthesis along with the other semantic elements). The device under consideration gets to its objective either by rearranging the groups emphasized or by inserting some emphatic particles (like именно ["just"], и ["too"], etc.). Besides, the said device rearranges the groups already located to produce an interrogative sentence.

The fourth device introduces into the half synthesized text pronoun substitutes, that is, replaces certain groups by the corresponding forms of the words он, она, оно, они—he, she, it, they. This operation may entail the change of position of the group replaced by the pronoun.

The fifth device evaluates the resulting sequence of definitive word groups from the point of view of a number of criteria (about 20) each of which singles out an undesirable property, or a construction to be avoided. Such properties are: (1) a very short group not carrying any emphasis is located nearer to the end of the sentence than a longer group; (2) there is no group to the left of the predicate group while to the right of it there are more than two groups; (3) there are three or more neighboring circonstants; (4) there is an ambiguity (the location of a group admits of two or more interpretations of its syntactic connection); (5) a circonstant separates the predicate group and one of the actants; and the like. Each of these properties is assigned a conventional "mark"—some negative number. Like arrangement weights, these numbers are chosen empirically. The output sentence is assigned a mark which is the sum of all marks associated with the undesirable properties and constructions in it.

Taking the sentence with the sum-total mark assigned to it, the fifth device tests in turn all possible rearrangements of the groups whose weight does not

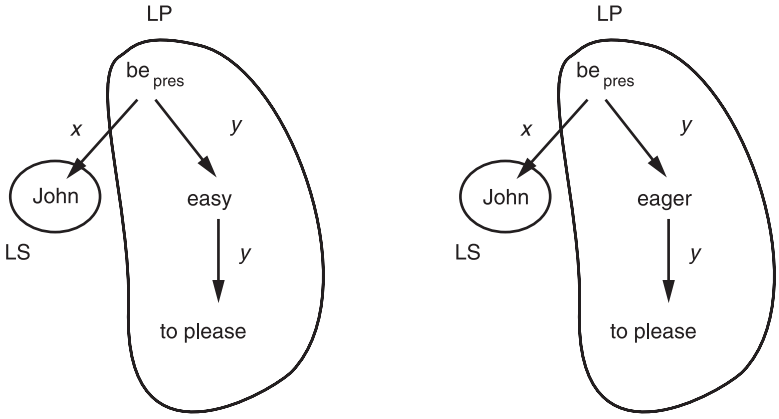


Figure 13.3
1_E) John is easy to please ⇒ (1_E)
2_E) John is eager to please ⇒ (2_E)

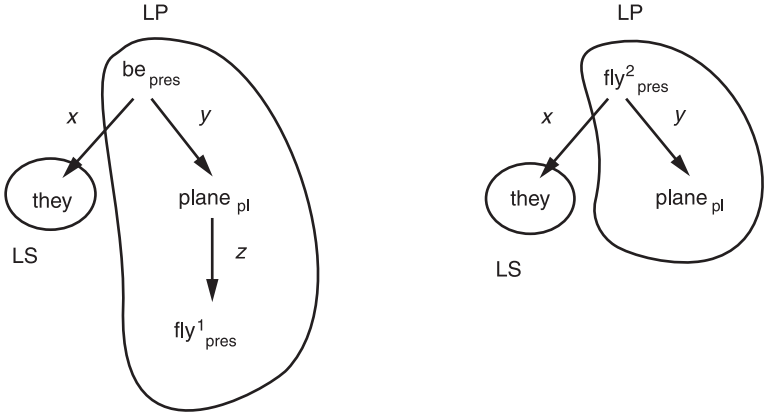


Figure 13.4
3_E) They are flying planes

exceed a fixed number, in order to get rid of as many as possible undesirable phenomena and thus raise the sum-total of the sentence. But it is not always possible to avoid all undesirable phenomena; a rearrangement cancelling one bad construction may entail another. Still, such a rearrangement may be helpful if the sentence's total rises; while admitting some undesirable phenomenon we eliminate a more undesirable one.

The result is either a sentence with the highest mark, zero (''sentence without drawbacks''), or a number of sentences with negative marks from which the one with the maximum mark (i.e. with minimum drawbacks) is chosen as the best one.

If, however, several sentences with equal marks are produced, the best one may be chosen by means of *preference rules*. Should these fail to select a single sentence as the best one, several variants considered equivalent in quality are sent to the output.

Failure by the algorithm to produce a zero mark output sentence means that the respective meaning cannot be expressed smoothly enough by means of what it has at its disposal. Some radical modification of the input syntactical tree (breaking it down into several independent sentence trees and/or replacing the vocabulary, i.e., the nodes, involved) may be needed. We are proposing to develop the sixth device capable of making such modifications.

Below (figures 13.3–13.14), an example of English-Russian (simulated) AT of some sentences ''via meaning'' is shown (with all necessary simplifications). The following abbreviations and symbols are used: 1_E), 2_E) ... —English sentences, 1_R), 2_R) ... —

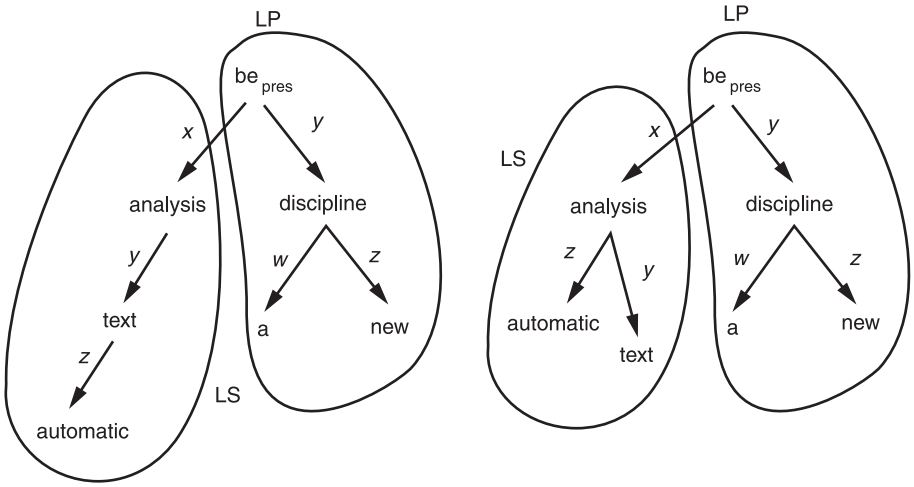


Figure 13.5
Automatic text analysis is a new discipline ($4'_E$) or ($4''_E$)

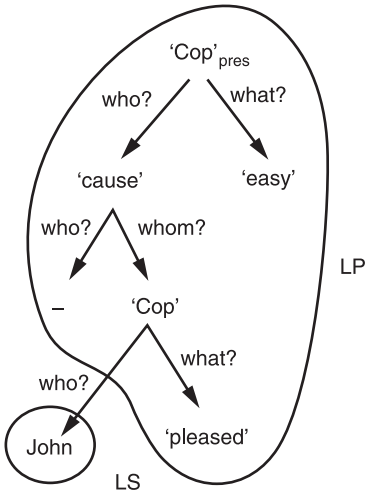


Figure 13.6
3. *Semantic analysis*

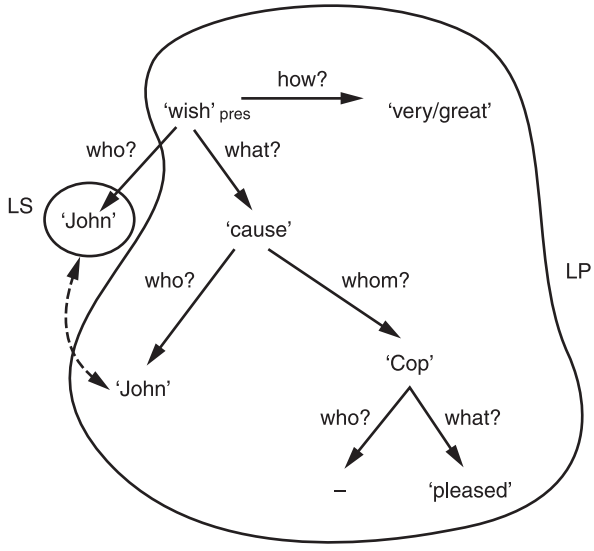


Figure 13.7
('To cause John be pleased is easy')

corresponding Russian sentences; 1_E) translates as 1_R) etc.; (1_E), (2_E) ... and (1_R), (2_R) ... —English and Russian syntactic trees, respectively; I, II, III ... — semantic trees; LS (and ЛС)—"logical subject" ("the known matter"); LP (and ЛП)—"logical predicate" ("the new matter"); words in quotation marks denote elements of semantic language; Cop (and СВ) = copula verb ("связка"); letters and figures marking the arrows of syntactic trees denote the type of syntactic connection (to denote the same connection type the same indices are used); the questions going with the arrows illustrate the type of semantic connection; dotted lines connect elements referring to the same extralinguistic object.

1–2. *Lexico-morphological and syntactic analysis*

($1'_R$) ⇒ $1'_R$) Джону легко доставить удовольствие.

($1''_R$) ⇒ $1''_R$) а. Джон—человек, которому нетрудно сделать прнятное.

б. Джон—такой человек, что ему нетрудно сделать приятное.

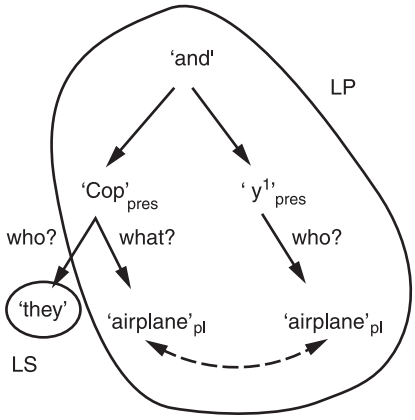


Figure 13.8
('John wishes very [much] that he (John) causes [someone] to be pleased')

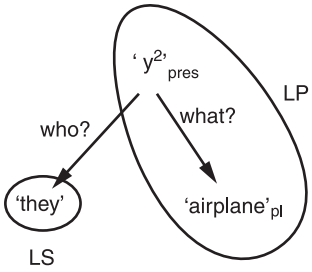


Figure 13.9
('They are airplanes, and [these] airplanes [are] flying')

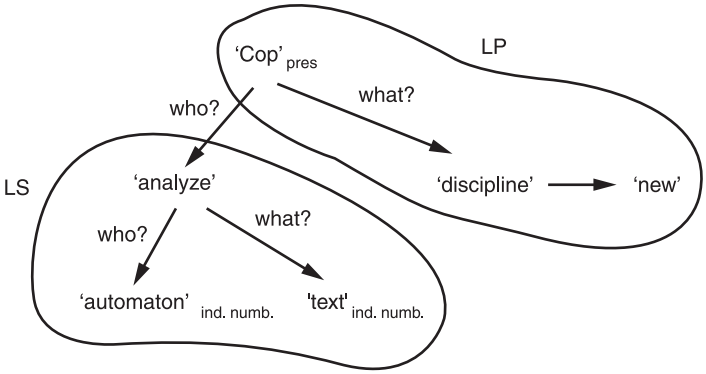


Figure 13.10
(4′): The syntactic tree is dropped by semantic analysis because of the semantic unacceptability of 'automatic text' (only devices, or actions and the like, can be 'automatic').

('Automaton[a] analyze[s] text[s]—is new discipline').

4. *Semantic synthesis*

$(2'_R) \Rightarrow 2'_R)$ Джон очень хочет доставлять удовольствие.

$(2''_R) \Rightarrow 2''_A)$ Джон жаждет делать приятное.

$(2'''_R) \Rightarrow 2'''_R)$ Джону очень хочется быть цриятным.

$(2''''_R) \Rightarrow 2''''_R)$ У Джона большое желание быть приятным.

$(3_R) \Rightarrow 3_R)$ Это летящие самолеты.

$(4'_R) \Rightarrow 4'_R)$ Они управляют самолетами.

$(4''_R) \Rightarrow 4''_R)$ Они ведут самолеты.

$(5'_R) \Rightarrow 5'_R)$ а. Анализ текста автоматами—зто новая дисциплина.

б. Анализ текстов автоматом является новой дисциплиной.

$(5''_R) \Rightarrow 5''_R)$ а. Автоматический анализ текста—новая дисциплина.

б. Автоматический анализ текстов представляет собой новую дисциплину.

Under synthesis, there are more possible syntactic trees for certain semantic structures, and more acceptable output sentences for certain syntactic trees, than we show here.

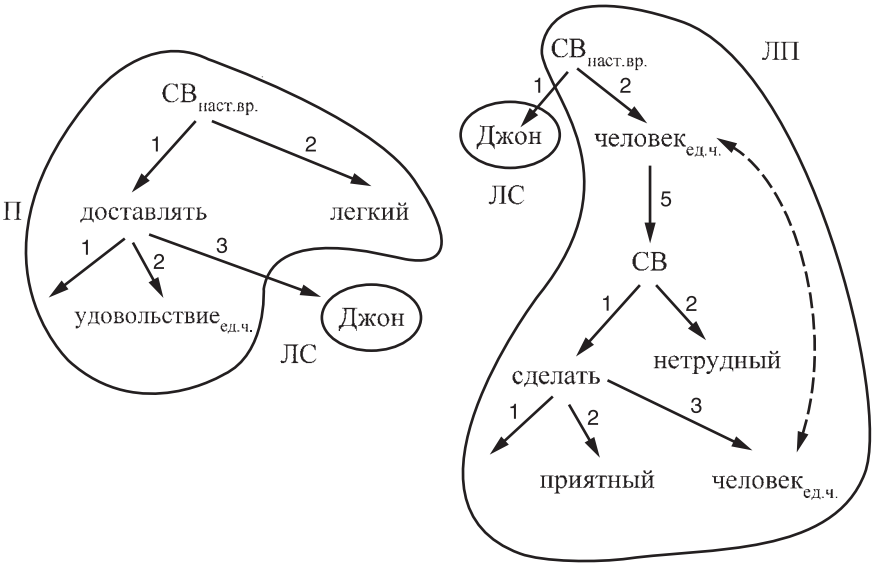


Figure 13.11

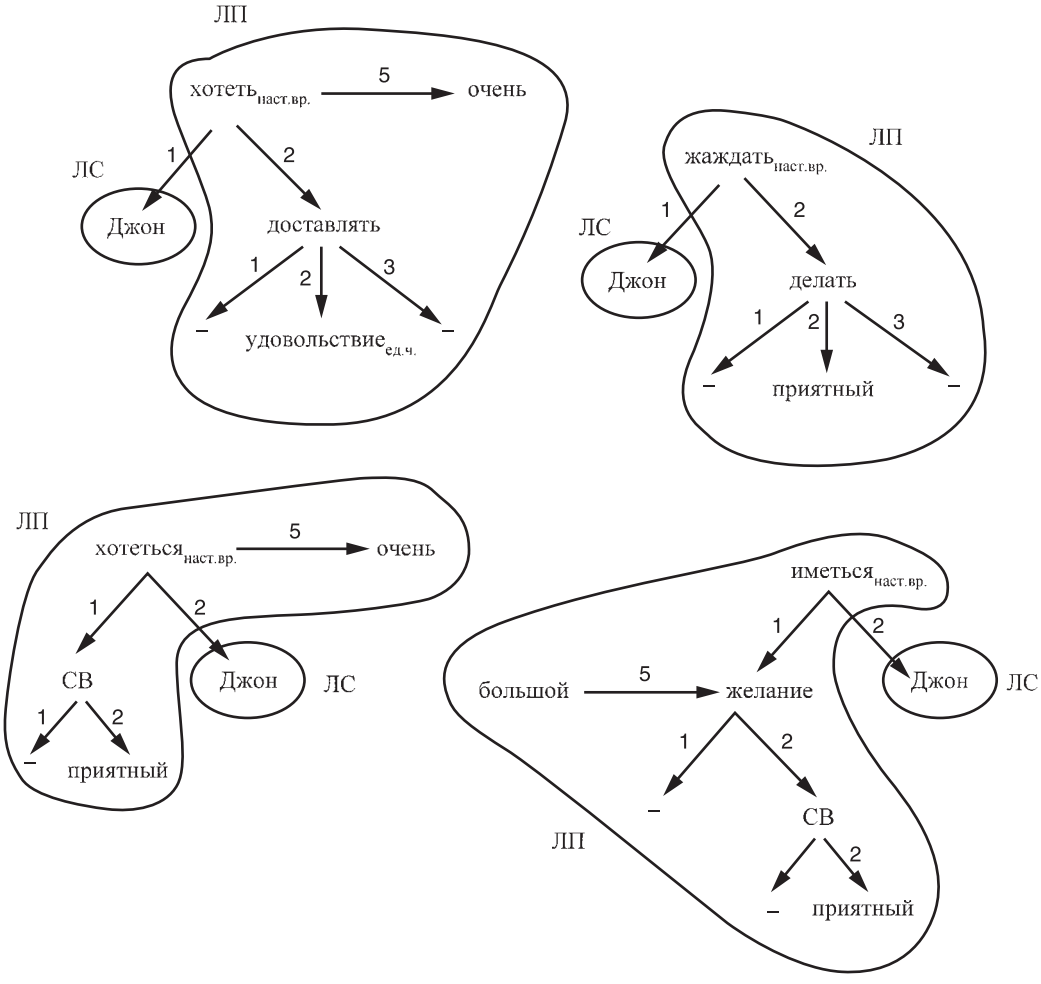


Figure 13.12

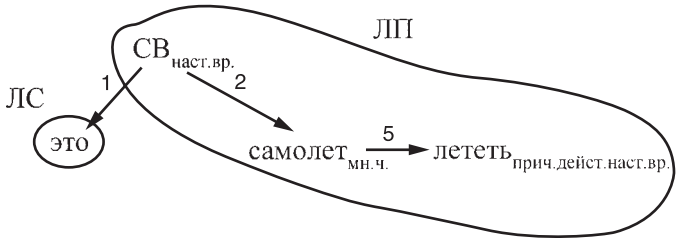


Figure 13.13

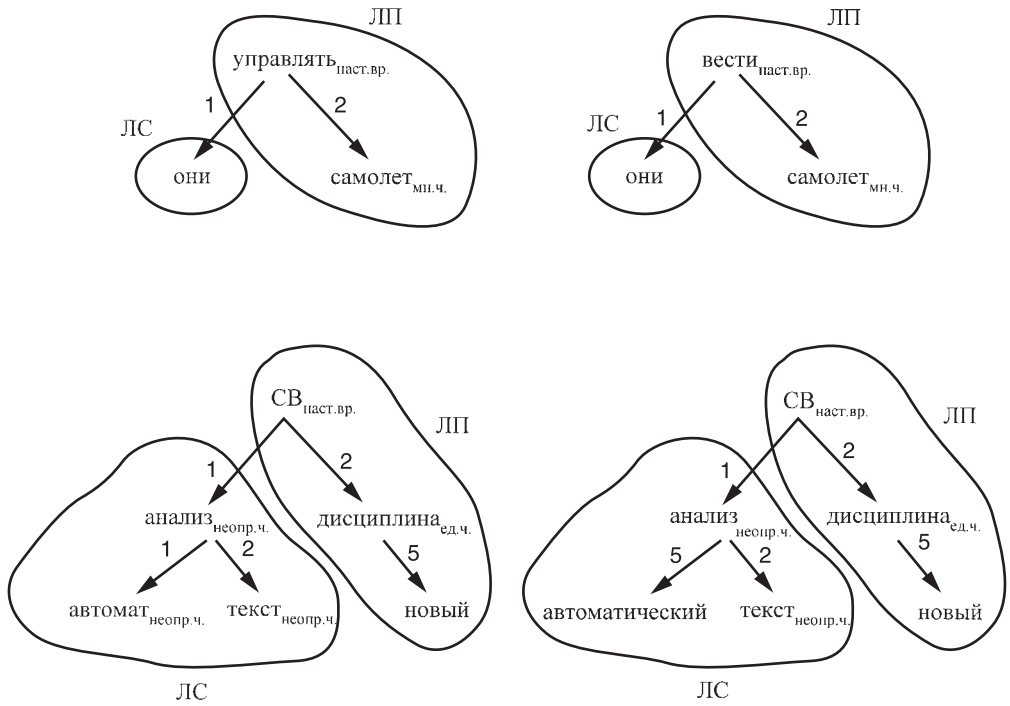


Figure 13.14

Acknowledgment

Contributors would like to acknowledge the valuable help of Victor H. Schnitke in preparing the English version of this chapter.

Notes

1. Our distinction "meaning/situation" corresponds to Frege's distinction "Sinn/Bedeutung".

2. Completeness and adequacy are understood here as stated in Mel'čuk [1963a].

3. The idea that "syntactic" translation is insufficient and that a special "semantic" stage is necessary has first been formulated in the USSR by the researchers of the Laboratory of MT in the Institute of Foreign Languages in Moscow (Žolkovsky et al. [1961], Mashinnyi Perevod [1964]). Their point of view and participation in their work have largely influenced the present writers.

4. If we agree to consider the variable (argument-name) as dependent on the corresponding predicate (function-word).

5. Here and further English words in quotation marks denote hypothetical (for the sake of illustration only!) elements of semantic language.

6. See footnote 4.

7. For types of syntactic connections see, e.g. Mel'čuk 1963*b*, 1964].

References

Alt, F. L., and I. Rhodes, 1962, Recognition of clauses and phrases in machine translation of languages, in *Proc. 1961 Intern. Conf. on*

Machine Translation of Languages and Applied Language Analysis (London).

Apresyan, Yu. D., 1963, Sovremennye metody izucheniya znachenii i nekotorye problemy strukturnoi lingvistiki, in *Problemy strukturnoi lingvistiki* (Moscow), pp. 102–105.

Beltrame, R., 1960, Illustration of the classification for developmental situations, *Methodos* (Milano) 12, no. 45–47, pp. 107–116.

Blois, J., E. Decresu, and J. Mommens, 1963, *Analyse morphologique automatique du français* (Bruxelles).

Bossert, W., 1960, The implementation of predicitive analysis, in *Mathematical linguistics and automatic translation* (Computation Laboratory of Harvard University), Rept. NSF-4, Cambridge, Mass.

Bossert, W., V. Giuliano, and S. B. Grant, 1960, Automatic syntactic analysis of English, in *Mathematical linguistics and automatic translation* (Computation Laboratory of Harvard University), Rept. NSF-4, Cambridge, Mass.

Ceccato, S., and B. Zonta, 1962, Human translation and translation by machine. I, in *Intern. Conf. on Machine Translation of Languages and Applied Language Analysis* (London).

C.N.R.S., 1961, *Les principes de l'analyse morphologique du russe dans le système de dictionnaire automatique*, Paris, Centre d'études pour la traduction automatique, Études préliminaires à la traduction automatique, vol. 7.

Corbé, M., and R. Tabory, 1962, Introduction to an automatic English syntax (by fragmentation), in *Proc. 1961 Intern. Conf. on Machine Translation of Languages and Applied Language Analysis* (London).

Dupuis, L., 1961, *Un système morphologique de la langue russe* (Centre national de la recherche scientifique. Centre d'études pour la traduction automatique.) No. 12.

Final Report, 1963, Linguistics Research Center, the Univ. of Texas, Rept. No. 16, (Austin).

Foust, W. D., 1960, Automatic English inflection, in *Mathematical linguistics and automatic translation* (Computation Laboratory of Harvard University), Rept. NSF-4, Cambridge, Mass.

Frege, G., 1892a, Über Sinn und Bedeutung, *Z. für Philos. und philos. Kritik* 100, pp. 25–50.

Frege, G., 1892b, Über Begriff und Gegenstand, *Vierteljahrschrift für Wissenschaft und Philosophie*, No. 16, pp. 192–205.

Garvin, P. L., 1961, Syntactic retrieval, in *Proc. of the National Symposium on Machine Translation*, pp. 286–292.

Giuliano, V., 1961, A formula finder for the automatic synthesis of translation system, *Mechanical Translation* 6, pp. 11–24.

Harper, K. E., and D. G. Hays, 1959, *The use of machines in the construction of a grammar and computer program for structural analysis*, RAND Corporation Rept. P-1624, Santa Monica, Calif.

Hays, D. G., 1961, Grouping and dependency theories, in *Proc. of the National Symposium on Machine Translation*, pp. 259–266.

Hays, D. G., and T. Ziehe, *Studies in machine translation. 10. Russian sentence-structure determination*. Santa Monica, The RAND Corporation (US Air Force. RM-2538).

Iordanskaya, L. N., 1964, Svoistva pravil'noj sintaksicheskoj struktury i agorithm ee obnaruzheniya (na materiale russkogo yazyka) in the series *Problemy kibernetike* 11 (Moscow), pp. 215–244.

Kulagina, O. S., 1962, Ob ispol'zovanii machiny pri sostavlenii algoritmov analiza teksta, in the series *Problemy kibernetiki* 7 (Moscow), pp. 209–233.

Kulagina, O. S., 1963, Ispol'zovanie mashin o issledovaniyakh po mashin o issledovaniyakh po mashinnomu perevodu, in the series *Problemy kibernetike* 10 (Moscow), pp. 205–213.

Lecerf, Y., 1961, Programme des conflits, modèle des conflits, *Traduction Automatiqaue* (Paris), no. 4, 11–18; no. 5, 17–36.

Lecerf, Y., 1961, L'adressage intrinsèque en traduction automatique, *Traduction Automatique* (Paris), no. 2–3, 31–47.

Mashinnyi, Perevod, *i prikladnaya lingvistika*, vol. 8, 1964, (Moscow).

Masterman, M., 1962, Semantic message detection for machine translation, using an interlingua, 1961, in *Proc. Intern. Conf. on Machine Translation of Languages and Applied Language Analysis* (London).

Mel'čuk, I. A., 1961, Morfologicheskii analiz pri mashinnom perevode (na materiale russkogo yazyka), in the series *Problemy kibernetike* 6 (Moscow), pp. 207–276.

Mel'čuk, I. A., 1963a, O standartnoi forme i kolichestvennykh kharakteristikakh nekotorykh lingvisticheskikh opisanii, in *Voprosy yazykoznaniya*, pp. 113–123.

Mel'čuk, I. A., 1963b, Avtomaticheskii analiz tekstov (na materiale russkogo yazyka), in *Slavyanskoe yazykoznanie* (Moscow), pp. 477–509.

Mel'čuk, I. A., 1964, *Avtomaticheskii sintaksicheskii analiz*, vol. 1 (Novosibirsk).

Mel'čuk, I. A., 1965, Poryadok slov pri avtomaticheskom sinteze russkogo texta (Predvaritel' noesoobshchenie), in *Nauchno-tekhnicheskaya informatsiya*, no. 12, pp. 36–41.

Moloshnaya, T. N., 1960, Algoritm perevoda s angliykogo yazyka na russkii, in the series *Problemy kibernetiki* 3 (Moscow), pp. 209–272.

Oettinger, A. G., 1960, *Automatic Language Translation* (Harvard Univ. Press, Cambridge, Mass).

Oettinger, A. G., 1961, A new theory of translation and its applications, in *Proc. of the National Symposium on Machine Translation*, pp. 363–366.

Oettinger, A. G., and M. Sherry, 1961, Current research on automatic translation at Harvard University and predictive syntactic analysis, in *Proc. of the National Symposium on Machine Translation*, pp. 173–182.

Paducheva, E. V., 1964, O sposobakh predstavleniya sintaksixcheskoi struktury predlozheniya, in *Voprosy yazykoznaniya*, no. 2, pp. 99–113.

Parker-Rhodes, A. F., 1961, Some recent work on thesauric and interlingual methods in machine translation, in *Information retrieval and machine translation*, part 2 (New York-London) pp. 923–934.

Progress Report No. 1, 1962, Nat. Sci. Found. on computer-aided research in machine translation, C 157-2 U 3 (Canoga Park).

Revzin, I. I., and V. Yu. Rozentsveig, 1963, Nekotorye voprosy teorii perevoda svyazannye s obshchei problemoi avtomatizatsii informatsionnykh protsessov, in *Nauchnotekhnicheskaya informatsiya*, no. 2, pp. 32–37.

Rhodes, I., 1959, *A new approach to the mechanical syntactic analysis of Russian*, Washington, 42, iv p. (National Bureau of Standards Rept. No. 6595).

Slutsker, G. S., 1963, Poluchenie vsekh dopustimykh variantov analiza teksta pri pomoshchi mashiny, in the series *Problemy kibernetiki* 10 (Moscow), pp. 215–225.

Tesnière, L., 1959, *Éléments de syntaxe structurale* (Paris).

Volotskaya, Z. M., 1961a, Formoobrazovanie pri sinteze russkikh slov, in *Lingvisticheskie issledovaniya po mashinnomu perevodu*, Soobshcheniya OMAIR, vol. 2 (Moscow), pp. 169–194.

Volotskaya, Z. M., 1961b, Voprosy slovoobrazovaniya pri mashinnom perevode, *ibid*., pp. 195–209.

Walking, J., 1960, The English Inflector, in *Mathematical linguistics and automatic translation* (Computation Laboratory of Harvard University), Rept. NSF-4, Cambridge, Mass.

Zasorina, L. N., N. B. Karachan, S. N. Medvedeva, and G. S. Tseitin, 1958, Proekt programm dlya morfologisheskovo analiza russkogo yazyka v mashinnom perevode, in *Materialy po mashinnomu perevodu*, coll. 1 (Leningrad).

Žolkovsky, A. K., N. N. Leont'eva, and Yu. S. Martem'yanov, 1961, O printsipial'nom ispol'zovanii smysla pri mashinnom perevode, in *Mashinnyi perevod*, vol. 2 (Moscow), pp. 17–46.

Žolokvsky, A. K., and I. A. Mel'čuk, 1965, O vozmozhnom metode i instrumentakh semanticheskogo sinteza, in *Nauchnotekhnicheskaya informatsiya*, no. 6, pp. 23–28.

This page intentionally left blank

14
Mechanical Pidgin Translation

Margaret Masterman

1 Introduction[1]

The basic problem in Machine Translation is that of multiple meaning, or *polysemy*. There are two lines of research which highlight this problem in that both set a low value on the information-carrying value of grammar and syntax, and a high one on the resolution of semantic ambiguity. These are (1) matching the main content-bearing words and phrases with a semantic thesaurus (Masterman [1956]) which determines their meanings in context; (2) word-for-word matching translation into a "pidgin-language" using a very large bilingual word-and-phrase dictionary. This paper examines the second.

The phrase "Mechanical Pidgin" was first used by R. H. Richens to describe the output given at the beginning of section 2 of this paper which, he said, was not English at all but a special language, with the vocabulary of English and a structure reminiscent of Chinese. Here it is used in a particular way, described below. Machine Translation output always is a pidgin, whose characteristics per se are never investigated. Either the samples of this pidgin are post-edited into fuller English; or the nature of the output is explained away as "low level Machine Translation"; or "rough Machine Translation" (Davies); or some vague remark is made to the effect that pidgin Machine Translation is all right for most purposes (International Business Machines [1959*a*]) (2). Thus, if a pidgin-dictionary is defined as one made by using the devices 1–4, given below, it might be said that the use of a pidgin-dictionary characterizes all Machine Translation programs. For in all programs a special dictionary is used to translate a limited subject matter; "pidgin-variables" (see below) form part of the output text; and some difficult grammatico-syntactic features of English (e.g., the use of certain auxiliary verbs, or of articles) are deliberately not accounted for by the program.

But there is a difference, indicated below by additional requirements.

For the Cambridge Language Research Unit we are deliberately setting out to accentuate and explore the 'pidginness' of pidgin as a language in its own right, on the assumption that it was a basic language.

The general requirements of a pidgin dictionary are the following:

1. Predominance of dictionary-entries for phrases rather than words.

2. Special sub-dictionaries, and the presupposition that a choice of subdictionary appropriate to the text has been made.

3. Specially constructed symbols here called "pidgin-variables," i.e., widely ambiguous words which the reader intuititively interprets according to the context (e.g., Reifler's he/she/it (Univ. of Washington [1958]) is a "pidgin-variable").

4. The omission of grammatical and syntactic features of the input language that a word-for-word Machine Translation program cannot transform.

The special requirements for a Mechanical Pidgin dictionary as defined here are the following:

5. It must not allow of any alternatives being included in the output, between which the reader of the output must find a way to choose. "The theory behind this rule is that a reader is less confused by a text containing occasional vague equivalents than by one containing all the possible equivalents of every word" (International Business Machines [1959*b*]).

6. The program must contain no provision for changing the word order of the text. This pinpoints the importance of studying what the older grammarians called "the actual sequence of ideas" (Allen and Greenough [1888]).

7. The pidgin must be treated and studied as a homogeneous language with properties of its own, *without consideration of the fact that different specimens of it may be derived from different source languages.*

The research went as follows: In 1959 a Latin–English Mechanical Pidgin dictionary of 700 entries was used as a control for other, more analytic, Machine Translation programs. The extreme difficulty

of doing better than the control stimulated interest in doing Mechanical Translation into pidgin for its own sake; and in November, 1959, an actual pidgin-producing machine program (for a punched-card laboratory) was constructed, debugged and operated. This program performed the same operations as the U.S.A.F.–I.B.M. photoscopic translation-system then performed, except that there was no "Rho-stuffing" program (International Business Machines ([1959*c*]). It chunked words into sub-words, not by a "peeling off" method (Reifler [1952]) but by a method called by R. M. Needham, who invented it, "exhaustive extraction" (Needham [1959], Kay and McKinnon-Wood [1960]). It had also a phrase finding procedure, and performed a one-one dictionary match. It had no device for changing word-order, nor for printing the output. Output from it is given in section 3 below.

In order to establish the notion of a Mechanical Pidgin, we start this paper with output obtained by Booth and Richens, and sophisticate this in stages, beginning with any two sentences, and using the four devices mentioned above. Section 2 is devoted to the construction of a pidgin dictionary for use in the program; and section 3 operates the program. Finally the potentialities of the work are estimated.

2 The Construction of a Pidgin Dictionary. Investigation of Booth and Richens's Mechanical Pidgin

2.1 The Text and the Pidgin Markers

The experimental material used was the Mechanical Pidgin output originally produced by Booth and Richens, and reported in Locke and Booth [1955].

Twenty sentences in different source languages were taken at random from the literature of plant genetics, sentences with proper names or numerical data being avoided. The samples were taken only from languages with Roman script, except for two sentences from oriental languages, Arabic and Japanese, which were transliterated to illustrate further points. In our investigations we treat these twenty sentences as if they came from a continuous text written in a single language. This continuous pidgin text is printed below, and is preceded by the list of pidgin markers used by Booth and Richens to indicate the function of words in the source languages. These markers, though capable of variable interpretation, are unambiguous in the sense that each marker is associated uniquely with a given class of inflections or constructions in any source language for which it is used. It is an additional assumption in all that follows that it is possible to define in use a single set of markers applicable to each of twenty different languages.

This output was not a Mechanical Pidgin, according to our definition, since many of the main words offered a choice of translations to the reader (these choices being separated in the output given below by a slash), and because the output contained no consciously contrived *pidgin variables*, that is, translations not occurring in English, designed to cover the whole range of meanings of a single word.

Pidgin-Markers of Richens's and Booth's Mechanical Pidgin

a	accusative	*o*	oblique
d	dative	*p*	past
f	future	*q*	passive
g	genitive	*r*	partitive
i	indicative	*s*	subjunctive
l	locative	*u*	untranslatable
m	multiple, plural, or dual	*v*	vacuous
n	nominative	*z*	unspecific

Pidgin Output of Richens and Booth in Continuous Form

S1. vine *z* enter in rest *z* in autumn/harvest *z* from/whence reason *zv* temperature *opv* low *z*.

S2. together work *z* between *m* country economic *z* union *m* and Danish *z* rural-dweller union *mg* seed/frog supply is continue *p* after same line *m* which/as in/you previous year.

S3. *v* disease come *z* thus very rapid up and has in many case *za*/one total amiss crop then *p* follow *z*.

S4. other *m* four foreign country (out of) standpoint *r*/standard *r*/bear are show oneself *pm* cultivation value *g*/*a* very insecure (become).

S5. *v* not is not/step astonish *v* of establish *v* that/which *v* hormone *m* of growth act *m* on certain species *m*, then that/which *v* not operate *m* on of other *m* if *v* one dream/consider *z* to *v* great *v* specificity of those substance *m*.

S6. in in a/one *d* large (more) area two form *m* beside one another live *z* without self to/too mix *z*, so belong/hear *pzz* different *m* form *m* circle *m* at.

S7. *v* small berry *v* variety *m* so crop/fruit quantity in as dry matter yield in surpass *mv* great berry *v* variety *ma*.

S8. (causative) sow *vm* sometimes thus enormous *v* damage *vv*, till/so that ought *v* sow once more/also.

S9. is been/status prove *p* that/which *v* cereal *m* of winter *z* grow *pm* in mountain crowd greenhouse show *m* little *v* resistance to cold while *v* same *m*/ is *ps* grown *pm* in field open *v* are much more resistant *m*.

S10. possible is however not prove/lacking *z* all *m* species/appearance same *g* genus/son-in-law *z* from same species/appearance *o* arise *z* draw *p*.

S11. however is able we already fixed *z* speak *v* concerning our river *z* oak forest *z* type *z* extensive (more) *z* spread *z* earlier *lm* time *lm* as also concerning this *z* that this *z* forest *z* dying out *d* at least through part *d* basis *l* been climatic reason *m*.

S12. growth *m* of autumn wheat was/wary more variable *m* from year to year than growth *m* or spring wheat.

S13. direction *m* bend *v* shot/trunk answer *m* direction *dm* dominant *om* wind *gm* and it behooves judge *v* that swordshaped (abstract noun) is cause *pq* through wind *m*.

S14. the/to existence of a/one number variable of seed *m* within of fruit show *z* that/which *v* various *m* ovule *m* of this plant has identical *v* possible (abstract noun) of self develop *v*.

S15. chromosome *m* barley *gm* cultivated *z* are of a/one diameter more great *z* than those *v* barley *gp* wild *z*.

S16. *v* study of *v* distribution of *v* temperature *m* minimum *m* annual *m* as is obvious *v* in all *v* work, reduce *vv* justification to *v* density of *v* station *m* and to record of observation *m* of each a/one of *v*.

S17. round/if earth *v* been freeze *p* long and deep, has no injury of clover rot *v* get *pq*.

S18. entire acidity view *p* (from) always rich is/ become (not) wine *m* our because malic acid decomposition condition *a*/*v* desire *d* suitable is not.

S19. and occur time *mz* division of chromosome *mv* limited *z* and that period *v* division *v* sperm *v* last result *zd* occurrence *m* mitotic *z*.

S20. this endure cold sex/disposition *g* difference (as for) tetraploid *n* diploid (at) sort/compare *d*/also osmotic pressure *n* high (adverb) becoming is fact *v* large *v* reason with/when consider *q*.

2.2 Sophistication of Two Sentences of the Text by Stages, to Form an Intelligible Pidgin Translation

The sentences chosen were those obtained from Italian and from Latin, i.e., S9 and S10.

The stages of sophistication were as follows:

Stage 1: Remove *z* and *v*.

To do this requires reference back to the two source languages, since it is often the case that Richens has thrown away information as *vacuous* and/or *unspecific* for a full English translation, which could be carried over into the pidgin by creating a pidgin variable.

The two original sentences were as follows (the asterisk * indicates a chunking-point):

S9. *Italian:* E stato prov*ato che i cerial*i d'invern*o cresc*iuti in serra mostr*ano proc*a resistenza a; freddo, mentre gli stessi cresc*iuti in campo apert*o sono molt*o piu resistent*i.

Output: *is* being/status prove *p* that/which *v* cereal *m* of winter *z* grow *pm* in mountain/crowd/ greenhouse show *m* little resistance to cold while *v* same *m*/is *ps* grown *pm* in field open *v* are much *v* more resistant *m*.

English: It has been proved that winter cereals grown under glass show little resistance to cold, while those grown in the open are much more resistant.

S10. *Latin:* Possibil*e est, at non expert*um, omn*es speci*es eiusdem generis ab eadem speci*e ort*um trax*isse.

Output: Possible *z* is however not prove/lacking *z* all *m* species/appearance same *g* genus/ son-in-law *z* from same species/appearance *o* arise *z* draw *p*.

English: It is possible, though not proved, that all species of the same genus have been derived from the same species.

Inspection of the above shows that Richens has translated by *z* (''unspecific'') all the Latin and Italian endings which are grammatically ambiguous; and he has translated by *v* (''vacuous'') all the Italian endings which are so very ambiguous that, in Richens' view, they mean nothing at all. As might have been expected, therefore, from the nature of the two languages, the pidgin output from Latin is sprinkled with *z*'s, but has no *v*'s, whereas the pidgin output from Italian is sprinkled with *v*'s but has only one *z*.

The *z* and *v* dictionary-entries which produced the two outputs, were as follows:

Latin		*Italian*	
-e	*z*	i	*v*
-um	*z*	-o	*z*
-is	*z*	-a	*v*
-adem	*z*	-e	*v*

Nothing can be done with these, pidginwise, so they are deleted from the output. The result is as follows:

From Latin:

possible is however not prove/lacking all *m* species/ appearance same *g* genus/son-in-law from same species/appearance *o* arise draw *p*.

From Italian:

is been/status prove *p* that/which cereal *m* of winter grow *pm* in mountain/crowd/greenhouse show *m* little resistance to cold while same *m*/is *ps* grown *pm* in field open are much more resistant *m*.

Stage 2: Convert *m*, *g*, *o*, *p*.
Refer back to the two source languages.

We will take the Latin sentence first. In the Latin, we come upon two mistakes:

1. *-es*, the ending which comes at the end of both *omn-es* and *speci-es*, is translated *m* in the first case, and not translated at all in the second case. Moreover, if the pidgin-dictionary is to have any generality *-es* cannot be translated *m*, since *-es* is also a 3rd declension singular ending; indeed it is the nominative singular ending of *species* itself. Even if Richens is picking up the '*m*' semantically, from the stem-meaning of *omn-is*, 'all', 'every', 'each', it becomes redundant if *omn-is* is translated 'all' and inappropriate if *omn-is* is translated 'each'. Thus the markers shown have been *z* in both cases, and hence should have been deleted at Stage 1.

2. *-e* cannot be translated by *o* when it occurs at the end of *speci-e* since it was translated by *z* when it occurred at the end of *possibil-e*. It should therefore be *z* in both cases, and hence should have been deleted at Stage 1.

g we pidginize as '-ish' for any language in which it comes out in the output post-positionally, and 'of+the' for any language in which it comes out in the output prepositionally[2].

p we pidginize as '-ed' for any language in which it comes post-positionally in the output, and 'did' for any language in which it comes pre-positionally in the output.

In the Italian case *-at*, translated by Richens *p*, we pidginize as '-ed' (see above), *-i*, which he translated by *m*, as '*-s*' (*on the assumption that the Plant Genetics pidgin-dictionary is never going to have any imperatives. N. B. This means making a special Italian pidgin-dictionary for cookery books*) and '*-ano*' we pidginize as '-they'.

The result is as follows.

From Latin:

possible is however not prove/lacking all species/ appearance same-ish genus/son-in-law from same species/appearance arise draw-ed.

From Italian:

is been/status prove-ed that/which cereals of winter grow-ed-s in mountain/crowd/greenhouse show-they little resistance to cold while same-s/is grow-ed in field open are much more resistant-s.

Stage 3: The creation of pidgin variables[3] (Rosten [1934]): '-ish' for *o* is already a pidgin variable.

Create other pidgin variables as follows:

Latin:	Italian:
'form' for *species*	(w) 'that' for *che*
'family' for *genus*	

The result is as follows:

From Latin:

possible is however not prove/lacking all form same-ish family from same form arise draw-ed.

From Italian:

is been/status prove-ed (w) that cereal-s of winter grow-ed in mountain/crowd/greenhouse show-they little resistance to cold while same -s/is grow-ed in field open are much more resistant-s.

Stage 4: The creation of phrases.

In the two sentences under discussion, the following phrases occur:

Latin:
1. *possible est* 'it+is+possible'
2. *non expertum* 'non+proven'
3. *ortum traxisse* '(to+have) derived+an origin'
4. *eiusdem* 'of+the+same'

Italian:
5. *e stato* 'has+been'
6. *cresciuti in serra* 'grown+under+glass'
7. *gli stessi* 'the+same'.

Final Result of Sophisticating the Translation of the Two Sentences. With the relevant phrases added to the dictionary, the Latin now becomes:

IT+IS+POSSIBLE HOWEVER NON+PROVED ALL FORM OF+THE+SAME FAMILY FROM SAME FORM (TO+HAVE)+DERIVED+AN+ ORIGIN.

The Italian now becomes:

HAS+BEEN PROVE-ED(W)THAT CEREAL-S OF WINTER GROWN+UNDER+GLASS SHOW-ED-THEY LITTLE RESISTANCE TO COLD WHILE THE+SAME GROW-ED IN FIELD OPEN ARE MUCH MORE RESISTANT-S.

Comment. Those to whom these sentences were shown had no difficulty in understanding them.

Stage 5: *Analysis of the whole text; frequency count and transition to the Interlingua 'Nude'.*

The following initial facts were obtained:

(i) Total number of pidgin words in the text 574

(ii) Number of sentences 20

(iii) Number of words in each sentence

(these are given together with a list of the source languages):

1)	Albanian	19
2)	Danish	30
3)	Dutch	22
4)	Finnish	20
5)	French	46
6)	German	32
7)	Hungarian	24
8)	Indonesian	18
9)	Italian	42
10)	Latin	22
11)	Latvian	54
12)	Norwegian	19
13)	Polish	31
14)	Portuguese	32
15)	Rumanian	22
16)	Spanish	43
17)	Swedish	19
18)	Turkish	23
19)	Arabic	28
20)	Japanese	28

A marker frequency-count was then made. The result of this is given below:

Marker frequency-count:
(the markers are arranged alphabetically)

a	3	*p*	14
d	7	*l*	3
g	7	*q*	3
m	53	*r*	2
n	2	*s*	1
o	3	*v*	51
		z	40

Total number of marker-occurrences 189

Total frequency-count of words in Richens' pidgin:

m	53
v	51
z	40
OF	20
p	14
IN; IN/YOU	10
IS	10
d; *g*	7
AND; TO; TO/TOO; THE/TO; MORE; (MORE); MORE/ALSO; NOT; NOT/STEP; NOT;	6
AS/ONE; AS; (AS FOR); WHICH/AS;	5
FROM; FROM/WHENCE; (FROM); SAME; THAT; THAT/WHICH; THIS;	4
l; *o*; *q*;	3
ARE; AT; (AT); BECOME; BECOMING; (BECOME);	
BEEN/STATUS; GREAT; GROWTH; HAS; REASON;	
SELF; ONESELF; SHOW; SPECIES/APPEARANCE; YEAR;	3
a; *n*;	2
(ABSTRACT NOUN); ALL; AUTUMN;	
AUTUMN/HARVEST; BARLEY; BERRY;	
(CAUSATIVE); CAUSE; CHROMOSOME;	
COLD; CONCERNING; COUNTRY; CROP;	
CROP/FRUIT; DIRECTION; DIVISION;	
CONSIDER; FOREST; FORM; FRUIT;	
HOWEVER; IF; LARGE; ON; ONE; OTHER;	
OUR; OUT; (OUT OF); POSSIBLE; PROVE;	
PROVE/LACKING; SEED; SEED/FROG; SO;	

SOW; TEMPERATURE; THAN; THEN;
THOSE;

THROUGH; THUS; TIME; UNION;
VARIABLE;

VARIETY; VERY; WHEAT;
WIND; WORD; 2

Alternatives (i.e., groups of words separated by an oblique stroke) are counted as single occurrences of each of the words. Thus CROP/FRUIT is counted as an occurrence of both CROP and FRUIT. The total number of occurrences is thus slightly higher than the number of words in the original text.

Inflected forms of the same root were counted as distinct words, e.g., DIFFERENT and DIFFERENCE were not counted as the same word.

Words in brackets were counted together with words not in brackets, e.g., (MORE) was counted together with MORE.

It is doubtful whether any inferences can legitimately be made from this frequency-count, given that all constituent sentences of the "text" came from different source-languages, and that the sample is such a small one.

If, however, we approach this "text" not linguistically but logically, in the older W. E. Johnson sense of "logic" as Universal Grammar (Johnson [1921]), it can be shown how this procedure, and others like it, suggested to R. H. Richens (the originator, with A. D. Booth, of Machine Translation from the British side) the first design of *Nude*, his interlingua (Johnson [1921], Richens [1956, 1959], Masterman [1962], Sparck Jones [1963]). And the same experiment can also be made to suggest a line-of-design, simpler than *Nude* and therefore easier to handle, for a Mechanical Pidgin to be used in word-for-word translation operating directly between two languages.

The graph (figure 14.1) shows the decrease of frequency of occurrence of the words in the sample when this is plotted against their rank order of occurrence. In his *Psychobiology of Language* Zipf discusses the relation between these quantities and concludes that it is in general of the form $ab^2 = k$. The graph we give was plotted for a very small sample of Machine Translation output and yet it does suggest that the Mechanical Pidgin displays this characteristic of natural language to which Zipf drew attention. On the basis of very large samples Zipf also came to the conclusion that this "law" is not obeyed by "... the few enormously frequent words ..." of a language (Zipf [1935]). This deviation for small values of rank

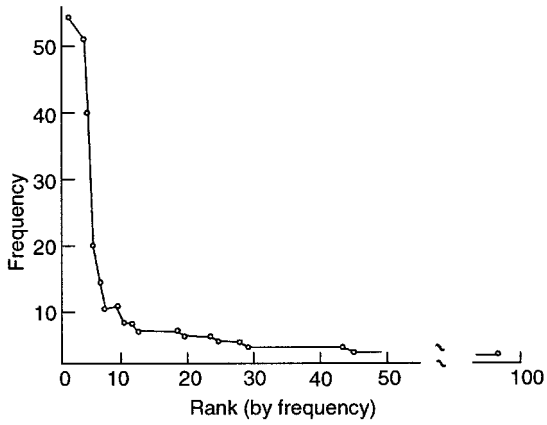


Figure 14.1
Richens and Booth's pidgin. Relative frequency of units plotted against rank by frequency (linear scale).

order is usually indicated by a small upward "bend" of the line in Zipf's graphs. Discussing this bending phenomenon (pp. 226 *et seq.*) Zipf makes the bold guess that the break in frequency of occurrence, indicated by this bend, represents a division of the words of a language into two groups. He suggests that these could function as the two fundamental parts of speech of the language.

If Zipf is right in this guess we may draw an inference as to the construction of a Mechanical Pidgin. Our sample is too small to show any deviation for the most frequent words. However, we may not only hope to find it in larger samples, we may also bias the construction of the Pidgin to accentuate any such deviation. We can do this by extending the use of "pidgin variables" to produce a class of words with a very wide range of meaning, and hence of very frequent occurrence. (In its untreated state the output's most frequent words were Richens's markers, including among them the vacuous and unspecific markers.) Thus we could reasonably expect to obtain a pidgin which reflected Zipf's distinction between Group I (the content words) and Group II (the deviants, the "bits and pieces of language") words. We would expect the Group II words to be predominantly pidgin variables. Each Group II pidgin variable would need to have both a pre- and post-positional form so that we could translate it optimally for English readers regardless of the nature of its representation in the source language.

2.3 Logical Analysis of Mechanical Pidgin

Acting on the above, we shall say that pidgin generated by an identical method from any input

language is a logically basic language, not an eccentricity. Using this assumption, it is easy to see how Richens came to choose the 50-odd elements of his interlingua, *Nude* (Richens [1956, 1959], Masterman [1962], Sparck Jones [1963]), from the first output, thus creating by his action a new, as yet undeveloped, field of semantic logic.

Comparison of the Group II Words, or Redefined Markers, Given by the Frequency-test, with the Basic Elements of Richens's Nude If we now redefine Richens' marker-list, regarding as arbitrary Richens' distinction between words indicated by lowercase letters and dictionary words, omitting *z* and *v*, and conflating *o* and *g*, we get the following reorganized marker-list in descending order of occurrence:

MANY; OF; PAST (i.e., *p*) IN; IS; THAT; AND; FOR (i.e., *d*); TO; MORE; ONE; NOT; THIS; AS; FROM; SAME;

If these then be compared by table with Richens' list of *Nude* elements, the semantic overlap becomes evident:

Table showing overlap between elements of Richens' Nude with the redefined set of markers:

Nude Elements	Markers	Nude Elements	Markers
NOT	NOT	WHERE	
DONE	PAST	BANG	
MUCH	MANY	WILL	
CAUSE		FOR	FOR
CAN		CHANGE	
LAUGH		WANT	
IN	IN	SENSE	
PRAY		HAVE	OF
DO		USE	
ASK		POINT	THIS, THAT
UP		SAME	SAME
FEEL		THINK	
MORE	MORE	BE	IS
COUNT		WHOLE	
TRUE		ONE	ONE
SELF		PLEASE	
FOLK		PART	
PLANT		MAN	
THING		BEAST	
WORLD		LINE	
LIFE		PAIR	
HEAT			STUFF
GRAIN			KIND
HOW	AS	WHEN	
	TO		FROM

2.4 Sophistication of the Whole Text

Using the techniques described above, we proceed to sophisticate the whole text given in section 2.1 with the aid of the information in tables 14.1 and 14.2.

Recognition and Translation of Phrases The designation of certain strings of words as "phrases" for the purpose of constructing a phrase-dictionary is closely

Table 14.1
Distinguishing Prepositional and Post-Positional Variants of Richens's Pidgin Markers (Sample Page)

Old form	New form	Position-in-text	Source language	Source-marker	Pidgin-translation
a2	-a	7,24	Hungarian	-at	(vacuous)
a2/v	-a	18,18	Turkish	-i	(vacuous)
d1	-d	6,4	German	-em	-WARD
d2	-d	11,40	Latvian	-ai	-WISE
d3	-d	11,45	Latvian	-a	-WISE
d4	-d	13,9	Polish	-om	-WISE
d5	-d	18,20	Turkish	-ya	-WISE
d6	d-	19,7	Arabic	1-	OF+THE
d7	d-	19,24c	Arabic	l1-	OF+THE
g1	-g	2,17	Danish	-ers	-WARD
g2/a	-g	4,7	Finnish	-n	-WARD
g3	-g	10,11	Latin	eiusdem	-ISH
g4	-g	13,15	Polish	-ow	-ISH
g5	-g	15,4	Rumanian	-ilor	-ISH
g6	-g	15,19	Rumanian	-ilor	-ISH
g7	-g	20,5	Japanese	-no	-ISH
11	-1	11,23	Latvian	-os	-POINT
12	-1	11,26	Latvian	-os	-POINT
13	-1	11,47	Latvian	-a	-POINT
m1	m-	2,5	Danish	de	SOME
m2	-m	2,10	Danish	-er	-S
m3	-m	2,16	Danish	-ers	-s
m4	-m	2,25	Danish	-r	-S
m5	-m	4,2	Finnish	Muut	-S
m6	-mA	4,14	Finnish	-neet	-THEY
m7	-m	5,15	French	-s	-S
m8	-mA	5,17	French	-issent	-THEY
m9	-m	5,20	French	-es	-S
m10	-m	5,22	French	-s	-S
m11	-m	5,28	French	-antes	(vacuous)
m12	-m	5,32	French	-s	-S
m13	-m	5,46	French	-en	-S
m14	-m	6,10	German	-en	-S
m15	-m	6,27	German	-en	-S
m16	-m	6,29	German	-en	-S
m18	-m	7,6	Hungarian	-ak	-S
m19	-mA	7,17	Hungarian	-ban	-THEY

Table 14.2
Showing Creation of Pidgin Variables Additional to Those Given in Table 14.1 (Sample Page)

Richens's output	Position-in-text	Pidgin variable	Comment
A/ONE	3;15;6;3;14;4;15;10;16;41;	AN(E)	
(ABSTRACT NOUN)	13;24;14;28;	-MENT	
(ADVERB)	20;18	-LY	
(AS FOR)	20;7;	-IN+REGARDING	
(AT)	20;11;	-THERE(AT)+TO	
AUTUMN/HARVEST	7;8;	AUTUMN+HARVEST (TIME)	
(BECOME)	4;20;	BECOMING	
BELONG/HEAR	6;22;	(AD)HE(A)R(E)	This stretches to the limit the device of making the context trick the readers' eye.
BEEN/STATUS	9;12;	STATUS	Picks up all forms of Italian verb 'to be' as phrases regardless of cost in phrase-increase.
(CAUSATIVE)	1,	CAUSE+	
CROP/FRUIT	7,8	CROP	
DO/ALSO	20,13	(MAKE)TOGETHER+WITH	
DREAM/CONSIDER 5,36	5,36	BROOD+ON	'dream' should not be in a plant genetics dictionary
FACT	20,21	THE+FACT+IS	i.e. 'FACT'
FROM/WHENCE	7,10	FROM+WHEREOF	
(FROM)	18,5	-FROM	
GENUS/SON-IN-LAW	10,12	FAMILY	'son-in-law' should not be in a plant genetics dictionary.
IN_7,IN_2,IN_4,IN_5,	1,4;1,7;3,11;6,2;	IN	
IN_8,IN_9,IN_{10}	9,15;9,33;16,18;		
IN_3/YOU	2,27	IN	Picks up Danish 'you' from verb.
IN_6,IN_7,	7,10;7,15;	IN REGARDING	

bound up with the decision as to what are the key "bits of information" that must be received if a text or its pidgin translation are to be understood. Thus the phrase-translation IT+IS+POSSIBLE of the Latin *possibile est*, when compared to the translation "IS POSSIBLE" obtained from the separate Latin words *possibile* and *est*, yields the bit of information that somebody is postulating something, and that what they are postulating is going to follow. Thus, the phrase tells one that "IS POSSIBLE" ought to be preceded not by "HE" or "SHE" but by "IT", so that it ought to be followed either by "THAT", making "IT+IS+POSSIBLE+THAT" or by "TO" making "IT+IS+POSSIBLE+TO". Syntactically, therefore, it tells one that a subsidiary sentence is contained in the main sentence. But this piece of information is not a "bit of information" in the general sense now being discussed, whereas the "bit of information" that someone (a human being) is postulating as possibly true what follows, is. (Note that "bit of information" is not being used here in the sense in which it is used in Information Theory.)

Provided this general "bit of information", which is part of the writer's argument, gets over somehow into the translation, it does not matter what specific English words, and what specific syntactic devices, are used to convey it.

This notion of "bit of information" helps to make a transition from the notion of a matching *dictionary* to that of a *thesaurus*; for phrases can be classified according to the "bit of information" which they convey. And such a "clustering" is the characteristic property of a thesaurus classification of the kind required for machine translation, as opposed to the vaguer forms of classification used in Roget. We can also draw the following conclusions on pidgin translation:

(i) An English pidgin designed as a language must have at least two parts of speech, i.e., content words, and the small set of frequently used variables.

Table 14.3
Technical Phrases

Technical phrase	Sequence of pidgin words which it replaces	Sentence number
BECOME+DORMANT	ENTER IN REST	S1
IN+AUTUMN	IN AUTUMN+HARVEST(TIME)	S1
CO-OPERATION	TOGETHER-WORK	S2
(STANDARD+VARIETY)	STANDARD+VARIETY	S4
GROWTH+HORMONES	THOSE(WHICH+ARE)HORMONES FOR+OF(M) GROWTH	S5
[SPECIFICITY]	SPECIFICITY	S5
FORM+CIRCLE	FORM-S CIRCLE	S6
GROWN+UNDER+GLASS	GROW-ED-S IN LUMP-HUMP	S9
IN+FORMER+TIMES	EARLIER-POINT-S TIME-POINT-S	S11
[AUTUMN+WHEAT]	AUTUMN WHEAT	S12
[SPRING WHEAT]	SPRING WHEAT	S12
[OVULE]	OVULE	S14
[CHROMOSOME]	CHROMOSOME	S15
MINIMAL+ANNUAL+TEMPERATURE	TEMPERATURE-S MINIMUM-S ANNUAL	S16
[MALIC+ACID+DECOMPOSITION]	MALIC ACID DECOMPOSITION	S18
CHROMOSOME+DIVISION	DIVISION OF+THE CHROMOSOME-S	S19
OF+SPERMATOGONIAL+DIVISION	WHICH+BE-DIVISION WHICH+BE-SPERM	S20
THE+OSMOTIC+PRESSURE	OSMOTIC PRESSURE-THAT+WE+ARE+TALKING+ABOUT	S20
[TETRAPLOID]	TETRAPLOID	S20
[DIPLOID]	DIPLOID	S20

(ii) Very large special word and phrase dictionaries will be needed for each special subject, e.g., 500,000 entries.

(iii) A thesaurus establishing synonym classes of words and phrases can be compiled if these can be classified according to the "bit of information" they convey.

Notes:

(i) It will be recalled that words that are chunked all in one piece count as a one word phrase.

(ii) Phrases given in square brackets in the above table are carried through unchanged from the raw output.

3 Control-translation into Mechanical Pidgin of a Portion of Caesar's *Gallic War*, Book I, as Generated by a Fully Mechanized Machine Translation Program

Since Caesar's *Gallic War* is famous as a text for translation, no English translation is appended. The pidgin generated was not further sophisticated but was left as it was.

3.1 Input Latin Text

Caesar—The Gallic War Apud Helvetios longe nobilissimus fuit et ditissimus Orgetorix. Is M. Messalla et M. Pisone consulibus regni cupiditate inductus coniurationem nobilitatis fecit et civitati persuasit, ut de finibus suis cum omnibus copiis exirent: perfacile esse, cum virtute omnibus praestarent, totius Galliae imperio potiri. Id hoc facilius eis persuasit, quod undique loci natura Helvetii continentur: una ex parte flumine Rheno latissimo atque altissimo, qui agrum Helvetium a Germanis dividit; altera ex parte monte Jura altissimo, qui est inter Sequanos et Helvetios; tertia lacu Lemanno et flumine Rhodano, qui provinciam nostram ab Helvetiis dividit. His rebus fiebat ut et minus late vagarentur et minus facile finitimis bellum inferre possent; quo ex parte homines bellandi cupidi magno dolore adficiebantur. Pro multitudine autem hominum et pro gloria belli atque fortitudinis angustos se fines habere arbitrabantur, qui in

Table 14.4
Unilingual Phrases: i.e., Phrases Which are Presumed to Justify Themselves in Use Because of Some Characteristic of the Source Language

Language-of-origin	Phrase	Pidgin	Sentence number
Albanian	BECAUSE+OF	FROM WHEREOF REASON	S1
Finnish	HAVE+SHOWN+THEMSELVES	ARE SHOW ONESELF-ED+BEEN-THEY	S4
French	IT+IS+NOT+ASTONISHING	THAT+(ONE)NOT-IS REALLY ASTONISH-ING	S5
French	TO+SUPPOSE	FOR OF (M) SUPPOSE-ING	S5
French	WHILE	WHEN(W)THAT	S5
French	OTHERS	FOR OF (M) OTHERS	S5
French	THINK+OF	BROOD(S) ON TO	S5
German	BELONG	(AD)HE(A)R(E) ... AT	S6
Italian	IT+HAS+BEEN+PROVED	IS STATUS PROVE-ED	S9
Latin	IT+IS+POSSIBLE	POSSIBLE IS	S10
Latvian	WE+CAN	IS-ABLE WE	S11
Latvian	CONCLUDING	FIXED SPEAK-ING	S11
Portuguese	INSIDE	WITHIN OF	S14
Spanish	IN+FACT	ON JUSTIFICATION	S16
Japanese	BY+COMPARISON	THERE+(AT)+TO	S20

longitudinum milia passuum CCXL, in latitudinum CLXXX patebant.

His rebus adducti et auctoritate Orgetorigis permoti constituerunt ea quae proficiscendum pertinerent comparare, iumentorum et carrorum quam maximum numerum coemere, sementes quam maximas facere, ut in itinere copia frumenti suppeteret, cum proximis civitatibus pacem et amicitiam confirmare. Ad eas res conficiendas biennium sibi satis esse duxerunt: in tertium annum profectionem lege confirmant. Ad eas res conficiendas Orgetorix deligitur. Is sibi legationem ad civitates suscepit. In eo itinere persuadet Castico, Catamantaloedis filio, Sequano, cuius pater regnum in Sequanis multos annos obtinuerat et a senatu populi Romani amicus appellatus erat, ut regnum in civitate sua occuparet, quod pater ante habuerat; itemque Dumnorigi Aeduo, fratri Diviciaci, qui eo tempore principatum in civitate obtinebat ac maxime plebi acceptus erat, ut idem conaretur persuadet eique filiam suam in matrimoniam dat. Perfacile factu esse illis probat conata perficere, propterea quod ipse suae civitatis imperium obtenturus esset; non esse dubium, quin totius Galliae plurimum Helvetii possent; se suis copiis suoque exercitu illis regna consiliaturum confirmat. Hac oratione adducti inter se fidem et iusiurandum dant, et regno occupato per tres potentissimos ac firmissimos populos totius Galliae sese potiri posse sperant.

3.2 *Chunked* (Kay and McKinnon-Wood [1960]) *Latin Text with Corresponding English Translations*; (Sample Page)

[COMPAR	GET+TOGETHER
ARE	-TO
[,	,
[I	—
[IUM	—
IUMENT	BEAST+OF+BURDEN
[UM	
[ENT	-THEY
[ORUM	-S OF
[OR	—
[UM	—
ET	AND
[CARR	CHARIOT
[OR	—
[ORUM	-S OF
[UM	—
[QUAM MAXIM	AS+MUCH+AS+POSSIBLE
[UM	—
[NUMBER	NUMBER+OF

Table 14.5
Pidgin Phrases

The following phrases, which also occur in the pidgin output, are produced by translating a single ambiguous word or chunk of the input language into a 'pidgin-variable' which consists of a whole pidgin phrase.

THAT+WE+ARE+TALKING+ABOUT
THAT+ONE+WHICH+IS
THOSE+WHICH+ARE
IN+ALL+ROUND
OF+THE
WHICH+BE

Pidgin translation after the inclusion phrases.

S1. VINE *THAT+WE+ARE+TALKING+ABOUT BECOME+DORMANT IN+*AUTUMN *BECAUSE+OF* TEMPERATURE-WISE *WHICH+BE* LOW.

S2. *CO-OPERATION* BETWEEN SOME COUNTRY ECONOMIC UNION-S AND DANISH RURAL-DWELLER UNION-S-WARD SEEN SUPPLY IS CONTINUE-ED BEEN AFTER SAME LINE-S (AS)+THAT IN PREVIOUS YEAR.

S3. *THAT+ONE+WHICH+IS* DISEASE COME-S THUS VERY RAPID UP AND HAS IN MANY CASE-S AN(E) TOTAL AMISS CROP THEN DID-FOLLOW.

S4. OTHER-S FOUR FOREIGN COUNTRY-*OUT*+OF STAND-POINT STANDARD+VARIETY. *HAVE+SHOWN+THEMSELVES* CULTIVATION VALUE-WARD VERY INSECURE-BECOMING.

S5. *IT+IS+NOT+ASTONISHING TO+SUPPOSE* (W)THAT *GROWTH+HORMONES* ACT-THEY ON CERTAIN-S SPECIE-S *WHILE* THEY ARE NONOPERATE-ING ON OTHERS-S, IF ONE *THINKS*+OF+*THAT+ONE+WHICH+IS* GREAT SPECIFICITY FOR+OF(M) THOSE SUBSTANCE-S.

S6. IF IN AN(E)-WARD LARGE-ER AREA TWO FORM-S BESIDE ONE ANOTHER LIVE WITHOUT SELF TO(O) MIX, SO *BELONG-THEY* DIFFERENT-S *FORM+CIRCLE-S.*

S7. THE SMALL BERRY-LIKE VARIETY-S SO CROP QUANTITY-IN-REGARDING, AS DRY MATTER YIELD+IN-REGARDING, SURPASS-THEY GREAT BERRY-LIKE VARIETY-S.

S8. CAUSE SOW-ING-S SOMETIMES THUS ENORMOUS BE DAMAGE-ED TILL SO OUGHT BE-SOW ONCE AGAIN.

S9. *IT+HAS+BEEN+PROVED* (W)THAT THE CEREAL-S OF WINTER *GROWN+UNDER+GLASS* SHOW-THEY LITTLE RESISTANCE TO COLD WHILE THE SAME GROW-ED-S IN FIELD OPEN ARE MUCH MORE RESISTANT-S.

S10. *IT+IS+POSSIBLE,* HOWEVER NOT PROVE, ALL FORM SAME-ISH FAMILY FROM SAME FORM ARISE DRAW-TO+HAVE.

S11. *HOWEVER WE*+CAN ALREADY CONCLUDING CONCERNING OUR RIVER OAK FOREST-S TYPE EXTENSIVE-ER-ISH SPREAD-THEY *IN+FORMER+TIMES,* AS ALSO CONCERNING THIS, THAT THIS FOREST DYING+OUT-WISE, AT LEAST THROUGH PART-WISE, BASIS-POINT BEEN CLIMATIC REASON-S.

S12. GROWTH-S OF AUTUMN+WHEAT WAS MORE VARIABLE-S FROM YEAR TO YEAR THAN GROWTH-S OF SPRING+WHEAT.

S13. DIRECTION-S BENDING-TRUNK, ANSWER-THEY DIRECTION-WISE-S DOMINANT-ISH-S WIND-ISH-S AND IT BEHOVES JUDS-ING THAT SWORD SHAPED-MENT IS CAUS-ED-BEEN THROUGH WIND-S.

S14. (TO)THE EXISTENCE OF AN(E) NUMBER VARIABLE OF SEED-S *INSIDE* FRUIT SHOW (W)THAT *THOSE+WHICH+ARE* VARIOUS-S OVULE-S OF THIS PLANT HAS IDENTICAL POSSIBLE-MENT OF SELF DEVELOPMENT.

S15. CHROMOSOME-S BARLEY-ISH-S CULTIVATED ARE OF AN(E) DIAMETER MORE GREAT THAN THOSE BARLEY-ISH-S WILD.

S16. THE STUDY OF *THAT+ONE+WHICH+IS* DISTRIBUTION OF *MINIMAL+ANNUAL+TEMPERATURE*, AS IS OBVIOUS IN ALL WORK, REDUCE *IN+FACT* TO *THAT+ONE+WHICH+IS* DENSITY OF *THOSE+WHICH+ARE* STATION-S, AND TO RECORD OF OBSERVATION-S OF EACH AN(E) OF THEM.

S17. *IF+ALL+ROUND* EARTH BEEN FREEZE-ED LONG AND DEEP, HAS NO INJURY OF CLOVER ROT GET-ED-BEEN.

S18. ENTIRE ACIDITY VIEW-FROM ALWAYS RICH IS+BECOME-NOT WINE-S THAT+WE+HAVE, BECAUSE MALIC+ACID+DECOMPOSITION CONDITION DESIRE-WISE SUITABLE-IS-NOT.

S19. AND OCCUR TIME-S *CHROMOSOME+DIVISION WHICH+BE*-LIMITED-IS(H), AND THAT PERIOD *OF+SPERMATOGONIAL+DIVISION WHICH+BE*-LAST, RESULT-IS(H) *OF+THE* OCCURENCE-S MITOTIC-IS(H).

S20. THIS ENDURE COLD (SEX) DISPOSITION-ISH DIFFERENCE-IN+REGARDING, AS FOR TETRAPLOID-*THAT+WE+ARE+TALKING+ABOUT* DIPLOID *BY+COMPARISON* (CON)SORT, UN+TOGETHER+WITH *THE+OSMOTIC+PRESSURE* HIGH-LY BECOMING-IS, THE+FACT+IS+WISE LARGELY REASON WHEN WITH CONSIDER-IS+BEEN.

[UM	—
[COEM	BUY UP
[ERE	—TO
,	,
[SEMENT	SOWING
[ES	—
QUAM MAXIM	AS+MUCH+AS+POSSIBLE
AS	—
[FAC	MAKE
[ERE	-TO
,	,
UT	THAT
IN ITINERE	ON+THE+JOURNEY
[COPI	RESOURCES
[A	—
[FRUMENT	CORN
[I	—
[SUPPET	BE+MADE+AVAILABLE
[ERET	-MIGHT
,	,
CUM	WHEN(ITH)
[PROXIM	THE+NEAREST
[IS	—

Note on the Symbolism of the Dictionary and Pidgin Output:

'+' connects words forming an output phrase.

'-' connects word stems and their appropriate output inflexions (see rule below).

'(' and ')' indicate a particular type of pidgin variable, as in the case of *(w)that* in section 2.2. In this output either these are variables ambiguous as to number, i.e., *parts(s)*, or are variables in meaning, e.g. *(ap)prove* or *when(ith)*. The last, for the Latin "cum", is to be understood as "when" or "with" depending on its context.

Rule Used to Decide Cases of Multiple Chunking

Many words in the preceding dictionary are chunked in a number of ways. "iumentorum", for example, has pidgin output for the following forms: [I, Ium, Iument, um, ent, orum, or, um]. If the outputs for all these were inserted in the pidgin translation of the passage, the reader would have to make choices as he read. Our word-for-word method precludes the examination of the context of "iumentorum" in order to determine a unique output, and hence the correct chunking of the Latin word. We therefore adopt the following rule to determine the correct chunking from the dictionary entry.

We assume that the chunking procedure allows us to distinguish "inflexion chunks" (those followed by a space in the text), from the others, which we call "stem chunks".

Rule: take the *longest* inflexion chunk ("i-chunks" for short), in the dictionary entry for the word. Write down the output for the corresponding stem chunk unless this is written *only* in chunked form, in which case repeat the procedure. Then write down the output for the longest inflexion chunk.

Example: "CUPIDITATE" has entries for CUPID, CUPIDIT, IT, AT, E. Longest i-chunk is "E", but corresponding s-chunk "CUPIDITAT" does not occur. Repeat for "AT": Corresponding s-chunk "CUPIDIT" exists chunked, *but also entire*. Thus write down outputs for "CUPIDIT", "AT" and "E".

3.3 Output English Translation

Among the+Swiss by+far noble-est was and rich-est the+chief+Orgetorix. He, during+the+consulate+of+M. Messalla and+M. Piso kingdom desire-'s induced conspiracy persuaded-s, that or limit-s own when(ith) all-s resources might+go+out-they: a+mere+nothing to+be when(ith) strength all-s excel-they+might, the+whole+of Gaul control to+gain+the+mastery+of. That+thing this the+more+easily to+them persuaded-s, in+respect+of+which on+all+sides the+nature+of+the+locality the+Swiss contain-they+are one however river the+Rhine wide-est and high-est, who district Switzerland from the+Germans divide-s; the+other however mountain(s) the+Jura high-est, who is between the+Seige-dwellers and the+Swiss; third Lake Leman and river the+Rhone, who province our from the+Swiss divide-s. From+all+this result-was that and less widely wander they+might+be+able; in+which+respect man war-to desirous great grief brought-they+were for-+the+sake+of honour war and bravery narrow-s self limit have-to declare-were. Who in length miles CCXL, in width CLXXX lie-they+were.

From+all+this were+led and by+the+authority+of the+chief+Orgetorix excited fixed-ad+they those+things+which to set+out-to tend-they+might get+together-to, beast+of burden-s+of and chariots+of as+much+as+possible number+of buy+up-to, sowing as+much+as+possible make-to, that on+the+

journey resources corn be+made+available-might. When(ith) the+nearest the+state-s peace and friendship confirm-to. To the+matter accomplish-to two+years self enough to+be considered-ed+they: in third year an+expedition by+law confirm-they. To the+matter accomplish-to the+chief+Orgetorix choose-is. He self deputation to the+state-s undertook-s. On+the+way persuade-s Casticus, the+chief+Catamantaloedes son, a+Seine-dweller, of+whom father kingdom in the+Seine-dwellers many-s year-s possessed-had and from the+senate of+the+Roman +people friend called there+was that kingdom in the+state-s own occupy-might, in+respect+of+which father before had-had; and+besides Dumnorix a+Haeduan, brother the+chief+ Diviciacus, who at+that+time the+predominant+influence in the+state-s obtain-was and+also mostly the+people acceptable there+was.

That the+same+thing attempt-might+be persuade-s and+they daughter own in+marriage gives. A+mere+nothing to+do to+be that (ap)prove-s attempt-s finish-to, because he+himself own the+state-s control obtain would might+be: not to+be doubt, but+that the+whole+of Gaul to+do+a+great+amount the+Swiss they+might+be+able; himself own resources and+own army that kingdom secured-would confirm-s by+this+speech+were+led from+one+another pledge and oath give, and kingdom occupied three powerful-est-s and+also strongest people-s the+whole+of Gaul self to+gain+in+the+mastery+of to+be+able hope-they.

"Garbage" Production Generated by Caesar Pidgin Dictionary In order to correct the misleadingly good impression conveyed to uninformed outsiders by the output given above, two sample "translations" of a passage from Newton's *Principia* and the first seven lines of Virgil's *Aeneid* follow. The former was chosen partly because it represents scientific writing in Latin, and also because the word order is far more closely related to English than classical Latin. It was hoped that this test would bring out the extent to which a word-for-word translation is affected by word order.

Latin Text 1: Newton's *Principia Mathematica*, Book I, Proposition LIX, Theorem XXII Corporum duorum *S* & *P*, circa commune gravitatis centrum *C* revolventium, tempus periodicum esse ad tempus periodicum corporis alterutrius *P*, circa alterum immotum *S* gyrantis, & figuris, quae corpora circum se mutuo describunt, figuram similem & aequalem describentis, in subduplicata ratione corporis alterius *S*, ad summam corporum S+*P*.

Pidgin Translation Body of+two *S* and *P*, about common centre+of+gravity *C* revolve-they, time periodic to+be to time periodic body one+or+the+other *P*, about the+other unmoved *S* circle-they, and form, which body about self mutually describe-they, form like and equal describe-they, in square+root the+reckoning body the+ other *S*, to whole body *S*+*P*.

Full Translation, for Reference, from Motte The periodic time of two bodies *S* and *P* revolving around their common centre of gravity *C*, is to the periodic time of one of the bodies *P* revolving round the other *S* remaining fixed, and describing a figure similar and equal to those which the bodies describe about each other, as $\sqrt{S}$ is to $\sqrt{(S+P)}$.

Note on the Experiment This experiment was carried out by hand; the Latin dictionary made for the Caesar text was used, with additions for the new words. Thus no attempt was made to construct a special dictionary; even when new words were added to the dictionary they were given as widely applicable translations as possible, for example, "form" for "figur-". The only exceptions were words which do not occur in classical Latin at all, such as "subduplicata" ("square root") and "gravitatis centrum" ("center of gravity").

Latin Text 2 Arma virumque cano, Troiae qui primus ab oris Italiam fato profugus Laviniaque venit litora, multum ille et terris iactatus at alto vi superum saevae memorum Iunonis ob iram, multa quoque et bello passus, dum conderet urbem inferretque deos Latio, genus unde Latinum Albanique patres atque altae moenia Romae.

Pidgin Translation Arms man-and sing, Troy who first from the+shore Italy destiny fugitive Lavinia-and come-s the+shore, much that on earth/terror tossing and high strength higher furious remembering Juno on+account+of rage, much also and war step/suffered/outspread, while found-might city bring+in-s and the+Gods Latin, race whence Latin Alba-and ancestors and high wall Rome.

Full Translation, for Reference, from the Edition of Page, Capps, Rouse, Warmington Post and Rushton Fairclough Arms I sing and the man who first from the coasts of Troy, exiled by fate, came to Italy and Lavinian shores; much buffetted on sea and land by violence from above, through cruel Juno's unforgiving wrath, and much enduring in war also, till he should build a city and bring his gods to Latium;

whence came the Latin race, the lords of Alba, and the walls of lofty Rome.

Margaret Masterman
C. L. R. U. July, 1960

4 After Five Years

By converting the program referred to in section 3 from punched-card form to computer form, more extended Mechanical Pidgin Translation experiments (Masterman and Kay [1960]) could obviously have been done. However, from the experiments we had done we considered that Mechanical Pidgin Translation had been tested to destruction.

The point of breakdown was this: semantic ambiguity can indeed be damped down by creating a very large number of particular phrases but these do not help the getting out of the generalized "bits of information" which make up the message.

To show this, imagine the length of these phrases progressively extended, to clause-length, sentence-length, paragraph-length, and finally text-length. With each extension the content will become more particularized, whereas what was needed, from the start, was to have it more general. Nor will unilingual syntactic analysis supply the right type of generality, though it may supply data for it; for, notoriously, the same "bit of information" can be differently expressed, both with regard to vocabulary, and with regard to syntax.

The way forward is:

(i) To accept the conclusion derivable from the Mechanical Pidgin Translation experiments that the phrase and not the word is the semantic unit of translation;

(ii) To make the machine cut the source text up into phrases (using syntactically and/or phonetically derived data), and then to do a dictionary match of these with a Mechanical Pidgin phrase dictionary in which classes of phrases are coded into sequences of pidgin-variables (e.g., into sequences of elements of Richens's *Nude* (Masterman [1961]).

As Alice found, in *Through the Looking-Glass and What Alice Found There*, after she had finished reading the poem Jabberwocky, the essential enterprise in deciphering a foreign text in an unknown language is to get hold of the "bits of information" of which the message basically consists.

Examples of such "bits of information" are: that a past action has occurred; that a comparison is being made between tetraploids and diploids with regard to the capacity of each to endure cold; that a statement is being made by somebody about something; that, as Alice said, "somebody killed something," and so on. These "bits of information" are more fundamental than the grammatical and syntactic features of the text;

(iii) To assign to these sequences of pidgin-variables a mathematically determinate recursive structure which can also be interpreted semantically as a Mechanical Pidgin structure (Wilks [1965*a*]). Thus the notion of a Mechanical Pidgin variable is abstracted from that of an English pidgin-variable; and the notion of the structure of a Mechanical Pidgin from that of a simplified English grammar and syntax;

(iv) To print, as first output, the structured concatenation of sequences of pidgin-variables: each such sequence conveying a "bit of information." This will be the message;

(v) To convert this output by some phrase-construction program into a sequence of phrases in the target-language.

This generalisation of the idea of a Mechanical Pidgin forms the basis of our present Machine Translation research program. Stage (v) has not been worked on as yet.

Notes

1. This paper has been condensed and revised from a workpaper with the same title, Cambridge Language Research Unit, M. L. 133, which was handed out at the U. S. Office of Naval Research Colloquium on Machine Translation held at the Princeton Inn, New Jersey, in July 1960 and which contains the relevant dictionaries in full. I am greatly indebted to D. S. Linney and Yorick Wilks for assistance in making the revision.

The original work was supported by the U.S. National Science Foundation, the U.S. Air Force Office of Scientific Research, and the U.S. Office of Naval Research, the revision by the Office of Naval Research.

2. This is the point where the Cambridge Language Research Unit parted company with the Reifler group. The Reifler group did not use the device of finding two different kinds of equivalents for 'grammatical' chunks, pre-positional and post-positional. The genitive, or possessive, case suggested only 'of' as an equivalent, and when it occurred post-positionally, it seemed to present an insuperable difficulty (see University of Washington [1958]).

3. Here the Cambridge Language Research Unit has merely carried further a tendency already noticed and remarked on by Reifler. "I knew, however, from my experience with a number of languages, that if mere 'accurate intelligibility' was wanted, one of three alternatives could very often represent all of them in all possible contexts. Thus we were faced with the task of making a wise selection of such representative alternatives." (See University of Washington [1958].)

References

Allen, J. H., and J. B. Greenough (eds.), 1888, *Caesar's Gallic War*, (London), pp. LVII–LVIII.

Davies, D. W., National Physical Laboratory, Teddington, personal communication.

International Business Machines, 1959*a*, *Final Report on Computer Set AN/GSQ-16* (*XW-1*), I. B. M. Research Center, Yorktown Heights, New York, vol. 1, pp. 12 and 20, and vol. 6, p. 13.

International Business Machines, 1959*b*, *Final Report on Computer Set AN/GSQ-16* (*XW-1*), vol. 6, p. 1.

International Business Machines, 1959*c*, *Final Report on Computer Set AN/GSQ-16* (*XW-1*), vol. 6, p. 55, App. C.

Johnson, W. E., 1921, *Logic* (Cambridge University Press, Cambridge), Part 1, p. xxi.

Kay, M., and R. McKinnon-Wood, 1960, *A Flexible Punched-Card Procedure for Word Decomposition*, M.L. 119, Cambridge Language Research Unit.

Locke, W. N., and A. D. Booth (eds.), 1955, *Machine Translation of Languages* (John Wiley & Sons, Inc., New York).

Masterman, M., 1956, *The Potentialities of a Mechanical Thesaurus*, M.L. 1, Cambridge Language Research Unit.

Masterman, M., 1959, *What is a Thesaurus?*, M.L. 90, Cambridge Language Research Unit.

Masterman, M., 1961, Translation, *Proceedings of the Aristotelian Society*, Supplementary Volume, pp. 169–216.

Masterman, M., 1962, Semantic Message Detection for Machine Translation, Using an Interlingua. *Proceedings of the International Conference on Machine Translation of Languages and Applied Language Analysis*, National Physical Laboratory, 1961 (Her Majesty's Stationery Office, London).

Masterman, M., and M. Kay, 1960, *Mechanical Pidgin Translation*, M. L. 133, Cambridge Language Research Unit, pp. 154 ff.

Needham, R. M., 1959, *The Problem of Chunking*, M. L. 75, Cambridge Language Research Unit.

Reifler, E., 1952, *The Mechanical Determination of the Constituents of German Substantiva Composita*. Studies in Mechanical Translation no. 7, Department of Far Eastern and Slavic Languages and Literature, University of Washington, Seattle.

Richens, R. H., 1959, Tigris and Euphrates, *Proceedings of the Symposium on the Mechanisation of Thought Process*, National Physical Laboratory, 1958 (Her Majesty's Stationery Office, London).

Richens, R. H., 1956, *A General Programme for Machine Translation Between Any Two Languages via an Algebraic Interlingua*, M. L. 5, Cambridge Language Research Unit.

Rosten, L., 1934, *The Education of Hyman Kaplan* (Gollancz, London), passim.

Sparck Jones, K., 1963, *A Note on Nude*, M.L. 164, Cambridge Language Research Unit.

University of Washington, Seattle, 1958, *Linguistic and Engineering Studies in the Automatic Translation of Scientific Russian into English*, AF30(602)-1566 and AF30(602)-1827.

Wilks, Y., 1965*a*, *Final Report*, AF 61-(052) 647, M.L. 171, Cambridge Language Research Unit.

Wilks, Y., 1965*b*, *Application of the C.L.R.U. Method of Semantic Analysis to Information Retrieval*, M.L. 173, Cambridge Language Research Unit.

Zipf, G. K., 1935, *The Psychobiology of Language* (Houghton Mifflin, Boston).

This page intentionally left blank

15
English–Japanese Machine Translation

S. Takahashi, H. Wada, R. Tadenuma, and S. Watanabe

1 Introduction

Machine translation has already been tried at several institutions and in most cases general purpose electronic computers with ample storage capacity have been used for this purpose. The only exception may be the machine of the University of Washington [1]. In Japan, the necessity of machine translation is probably more intense than in other countries, because Japanese people have particular difficulties in learning foreign languages, due to their quite different letters and to the unique grammatical structure of their language.

The usual difficulties of machine translation are also found in the programming for translation from English to Japanese. It would be better, therefore, to examine translation principles thoroughly, using a general purpose computer, before constructing a special purpose machine. In Japan, however, computer development is still in an early stage, and until recently only a few computers, all with a storage capacity less than 1,000 words, were available. Therefore, it was decided to construct a special purpose machine, which has a relatively large magnetic drum store and handles words of variable length, but which has neither multiplication nor division mechanisms. This machine was completed about six months ago and named "Yamato," which meant "Japan" in ancient times.

At the same time, the textbooks on English which were being used in the first and second grade classes of some Japanese junior high schools were investigated. 2,000 English words were picked out, and a program flow chart to translate the whole textbook of the first year grade was prepared. At this stage neither relative pronouns nor relative adverbs appear and present perfect tense has not yet been used, but this would be the pertinent stage for the first trial of English-Japanese machine translation. A test of the program on Yamato has now been conducted. This paper describes translation principles and program flow diagrams at this stage as well as the organization of Yamato.

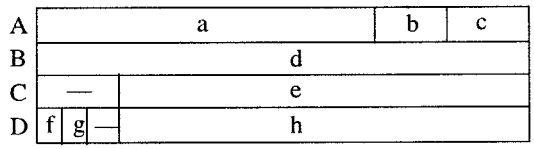
2 Dictionaries and Tables

Four kinds of dictionary: word, idiom, syntax, and Japanese word dictionaries are stored on the magnetic drum. In the dictionaries the length of each separate item, an idiom for instance, is naturally variable. Three kinds of table, in which each item is a fixed length code word of eight characters each of eight bits, are also stored on the drum. These correspond to the word, idiom and syntax dictionaries, and are called the word, idiom and syntax tables respectively.

2.1

The word dictionary is simply an arrangement of 2,000 English words in the order of probable frequency of use. The word table, on the other hand, is an arrangement of eight character code-words which correspond to the contents of the word dictionary, one by one and in the same order. Each code-word indicates the grammatical features of the corresponding English and Japanese words, the location of the latter in the Japanese word dictionary, etc. By consulting the word dictionary, each word of the given English text is transformed into the address of the corresponding eight characters in the word table.

Figure 15.1 indicates the structure of such an eight character code-word. In the figure A, B, C and D each consists of two characters of eight bits, one bit of which is a parity check bit and has been omitted for simplicity of explanation. In the eight character code-word there are a number of separate items indicated by a, b, etc.; a, d and e indicate the English part of speech, the Japanese part of speech and the location of the corresponding Japanese word, respectively; b and c are the locations reserved for the information concerning the affixes which are removed from the original word when the word dictionary is consulted, C indicates the existence of an affix and b its type. The bit f denotes whether the word occurs with other items in the word dictionary in at least one idiom. The bit g denotes whether the word has another meaning of different part of speech or not and h gives the



The first meaning of "spring"

A	0	0	0	0	1	1	0	0	0	0	0	0	0	0
B	1	0	0	0	1	1	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	1	1	1	0	1	1	1	1
D	0	1	0	1	0	1	0	0	1	0	1	1	0	0

The second meaning of "spring"

A	0	0	1	1	1	1	0	0	0	0	0	0	0	0
B	0	0	0	0	0	0	0	0	0	0	0	0	0	0
C	0	0	0	0	1	0	0	1	1	0	1	0	1	0
D	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 15.1
Contents of eight character code-word with an example. A, B, C and D are two characters each.

location of another eight character word corresponding to this meaning in the word table.

In figure 15.1 the word "spring" is shown as an example. Two meanings of "spring" appeared in the textbooks investigated; one is the intransitive verb which means トビハネル (tobihaneru), that is "leap" or "jump," and the other is the noun which means ハル (haru), "this season of the year." Fortunately the noun which means バネ (bane), "un ressort" in French, has not appeared. At the present stage multiple meanings for the same part of speech must be avoided since the program storage capacity is limited.

For the first meaning in the example, the two characters represented as A indicate that it is a root of a perfect intransitive verb; B, that its Japanese equivalent is a verb which is conjugated following a rule indicated; C, that its Japanese equivalent is the 751st word in the Japanese dictionary; D, that it does not occur in any idiom, but that it has at least one other meaning, which is given in the 1324th location of the word table.

For the second meaning, A denotes that it is a common noun; B and C, that its Japanese equivalent is a common noun and the 618th word in the Japanese dictionary; D, that it has no further meaning.

2.2
The idiom dictionary is an arrangement of groups of words, also in the order of frequency of use. Each word is represented by the two characters D denoting the location of any other meaning. For the word which has no other meaning but occurs in one or more idioms, a number denoting a pseudo-location is given. Whenever more than two words having "I" in the bit f appear successively, this dictionary is consulted. The biggest group of words which coincides with a content of the dictionary is assumed to be an idiom in the given sentence, and is changed into the address for the corresponding eight character word in the idiom table. Idioms which consist of words separated from each other, such as "so ... that ..." are excluded from this dictionary and must be treated by a program.

2.3
The syntax dictionary is an arrangement of 20 groups of parts of speech, corresponding to the part A of the code-words. The syntax table consists of addresses indicating the beginning of the program subroutine that should be used for the corresponding syntax.

2.4
The Japanese word dictionary is simply an arrangement of Japanese words.

3 Translation Principles and Flow Diagram

Figure 15.2 shows in outline the principles of English-Japanese machine translation. A given sentence is read in word by word, and changed into a series of eight character code-words with the use of the word dictionary and table. Any word which cannot be found in the word dictionary, even after the removal of affixes, is put in a separate track of the magnetic drum for the purpose of direct type-out (transliteration). It is assumed to be a noun, and is also replaced by an eight character code-word. The idiom dictionary is used after a whole sentence has been read in.

Since the grammatical structures of the English and Japanese languages are quite different, "word for word" translation, such as might be useful for translating between two European languages, is generally unsuccessful. It is important to find the grammatical structure of a given English sentence first, and then to transpose the words according to the corresponding Japanese grammar. The syntax dictionary is consulted for this purpose.

Before consulting the syntax dictionary, however, it is necessary to simplify the given sentence to the form of a basic pattern which is included in the dictionary; "noun+transitive verb+noun," for instance. The procedure for doing this will be described later. If

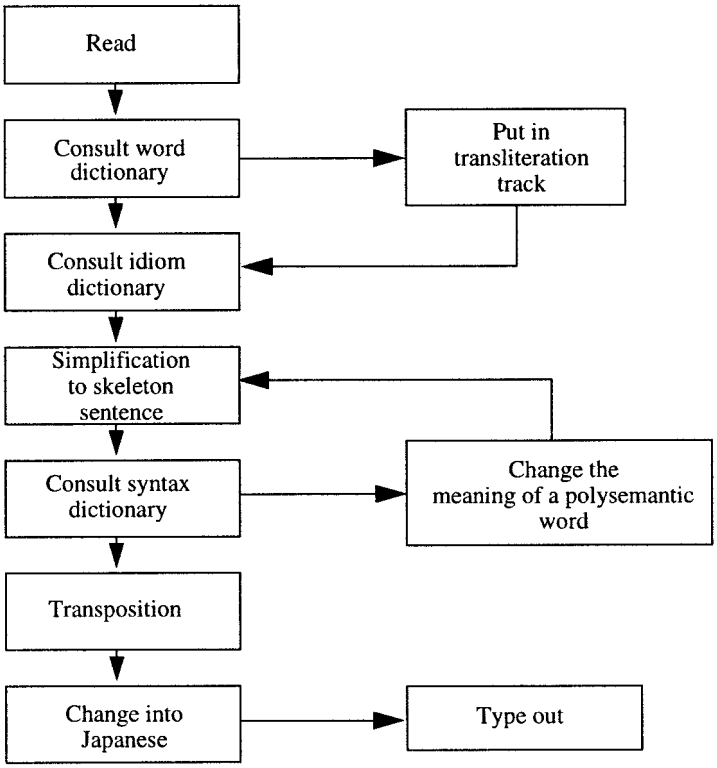


Figure 15.2
Translation principle.

the simplified pattern for the given sentence is found in the syntax dictionary, then the subroutine for the word transposition is given by the syntax table, as stated in section 2.3. If it is not found, the meanings of the polysemantic words are changed one at a time, and the simplification processes are repeated from the beginning. The multiple meanings are arranged in a sequence which considers each word first as an auxiliary verb and finally as a noun, in order to prevent endless iterations. If the pattern cannot be found by any means, the sentence is translated word for word.

The subroutines not only transpose the words, but also insert certain words which are peculiar to the Japanese language. After this, every group of words is decomposed into the original eight character codewords, and each word is changed into Japanese with the use of the Japanese word dictionary. Finally the words are typed out one by one. Japanese sentences are generally written in a mixture of three kinds of letter; "kanji" (Chinese ideographs), "hirakana" and "katakana." In the present experiment only "katakana," consisting of 75 letters is used. The alphabet (capital letters only) also can be typed out, and is used for transliteration.

A very simple example of the translation processes is shown in figure 15.3, and the program flow diagram in figure 15.4. It takes about ten seconds to complete the translation of figure 15.3, including the time required for the input and output. The whole translation program of figure 15.4 requires about 4,000 instructions.

The process of simplification to the basic pattern is shown in figure 15.4. The process may be divided into three parts: 1) grouping of words, 2) decreasing the number of verbs to one, 3) removing adverbs. The last part is simple and requires no explanation. The second part is described fairly precisely in figure 15.4. The first part is subdivided into five types of grouping; (a) auxiliary verb+verb verb, (b) adverb+adjective adjective, (c) adjective or possessive form+noun noun, (d) article+noun noun, and (e) preposition+noun adverb. In the case of the preposition "of," the group (e) often reduces to an adjective modifying the preceding noun. Therefore, "of" is treated separately and the preceding word is checked.

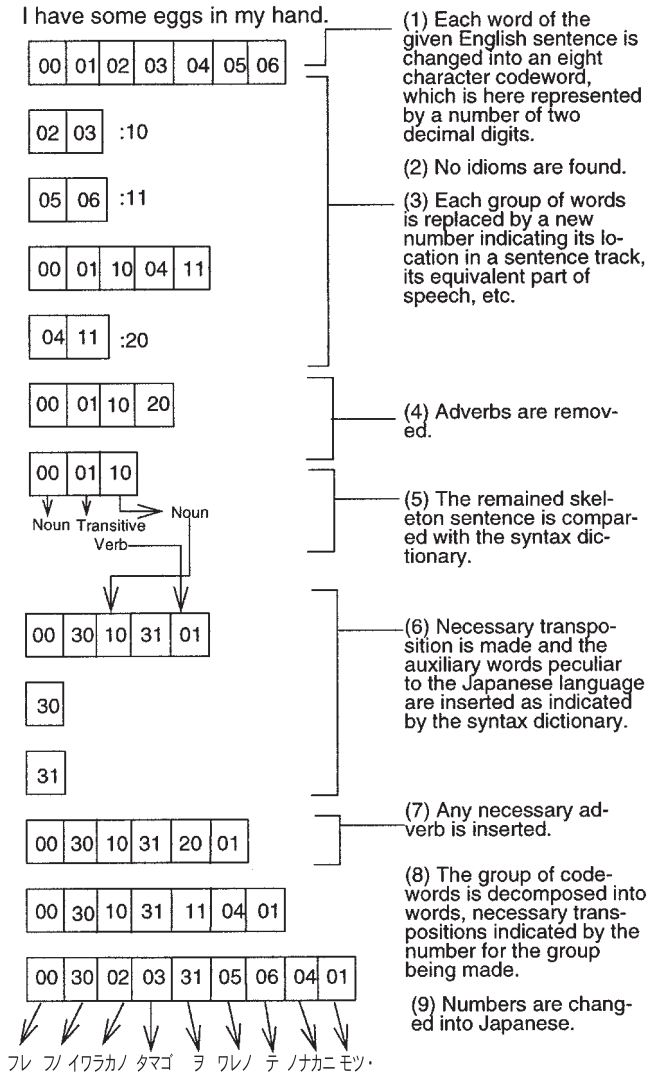


Figure 15.3
A simple example of English–Japanese machine translation.

4 Machine Organization

Yamato is a binary serial computer and operates at a clock repetition rate of 195 kc. Instructions and numerical words are both represented by 32 bits, including one bit for a parity check. A magnetic drum of 200 tracks, having a capacity of 820,000 bits and an average access time of 10 milliseconds, is used as the store. A one-plus-one address code is employed to compensate for the slow speed of the drum. The address parts of the instruction are 12 bits each, and 6 bits are used for the functional part.

Yamato can execute 46 different instructions, detailed in table 15.1. Each instruction has two address parts, A_1 and A_2. The location of the next instruction is always indicated in A_2, except in the case of the jump instruction. For some instructions A_1 indicates an address in the program section of the store while for others it supplements the functional part in the designation of the operation. In the latter case, A_1 is subdivided into three hexadecimal numbers A_{11}, A_{12}, A_{13}.

Since Yamato is a special purpose computer, some of the instructions are peculiar to it. It handles information of variable length, such as an English word which may have up to 16 characters for which two letter registers of 8 characters each are used for this purpose. There are also four counters to count word or letter numbers automatically in some operations.

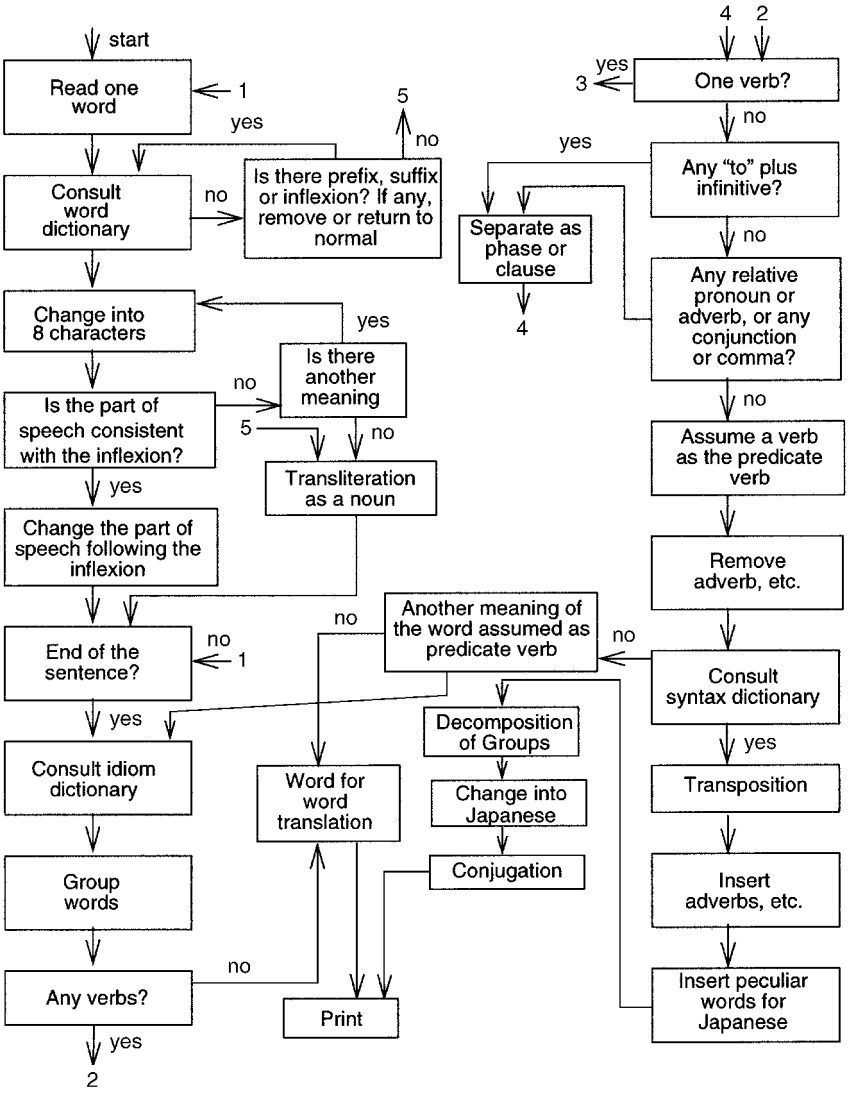


Figure 15.4
Flow diagram.

The storage is divided into three large sections: the dictionary, table and program section. The dictionary section, which includes the tracks for the words to be transliterated, stores information of variable length. Neighboring words are separated from each other by an "all mark," namely a character whose eight bits are all 1. When a word is stored in this section, the "all mark" is automatically inserted directly behind the word. The address of a word in this section is determined by counting the "all" marks from the top. The table section, including the tracks which store the sentence in various forms as it is treated, stores only eight character code-words. The program section has 4096 locations and stores instructions and numbers of 32 bits.

The English text is presented in the form of punched tape. This is punched by hand at present, but in the near future the punching will be performed automatically by a reading machine [2].

Being completely transistorized, Yamato consumes only 50 watts excluding the power for the drum motor and for the input and output. It employs dynamic basic circuitry, including a one bit delay which was invented by Takahashi, one of the authors, and which has been successfully operating in a general purpose computer, ETL Mark IV [3], for more than a year.

Table 15.1
Instructions of Yamato

Code	Abbreviation	Operation
02	Add	The number in A_1 is added to the accumulator
03	Clear Add	The number in A_1 replaces the contents of the accumulator
04	Sub	The number in A_1 is subtracted from the accumulator
05	Clear Sub	The negative of the number in A_1 replaces the contents of the accumulator
06	Store	The contents of the accumulator are copied to A_1
07	Clear Store	Both the accumulator and A_1 are cleared
10	Read in to Acc	One character on the tape is read to the accumulator. The previous contents of the accumulator are shifted A_{13} places to the left
11	Clear Read in to Acc	The accumulator is cleared. The other actions are the same as in 10
12	Read in to LR	One English word, not longer than 16 characters, is read to the letter register
14	Acc Shift	The contents of the accumulator are shifted A_{13} places; to the left when A_{11} is even, and to the right when A_{11} is odd
20	Raise Acc	A_1 is added to accumulator
21	Clear Raise Acc	A_1 replaces the contents of the accumulator
22	Lower Acc	A_1 is subtracted from the accumulator
23	Clear Lower Acc	A_1 replaces the contents of the accumulator
24	Add Counter	The contents of the counter designated by A_{13} are added to the accumulator
25	Clear Add Counter	The contents of the counter replaces the contents of the accumulator.
26	Acc to Counter	A part of the accumulator contents replaces the contents of the counter designated by A_{13}
27	Clear Acc to Counter	Both the counter designated by A_{13} and the accumulator are cleared
30	Add Letter	The A_{11}th letter in the letter register is added to the accumulator
31	Clear Add Letter	The A_{11}th letter replaces the contents of the accumulator
32	Acc to LR	The 7 least significant bits of the accumulator replace the A_{11}th letter of the letter register
33	Clear Acc to LR	Both the accumulator and the A_{11}th location of the letter register are all cleared
34	Extract to Acc	The logical product of the 7 least significant bits of A_1 and the A_{11}th letter of the letter register is brought to the accumulator. The previous content of the accumulator is shifted 7 places to the left
35	Clear Extract to Acc	The accumulator is cleared before the logical product stated above is entered
40	Bring from Table to LR 1	An eight character code-word is brought to the letter register 1 from the location in the table section of the store which is defined by A_{11}, A_{12} and the contents of the counter designated by A_{13}
41	Bring from Table to LR 2	An eight character code-word is brought to the letter register 2 from the location defined in 40
42	Store LR 1 to Table	The contents of the letter register 1 are copied to the location defined in 40
43	Store LR 2 to Table	The contents of the letter register 2 are copied to the location defined in 40
50	Bring from Dictionary	A word is brought to the letter register from the location of the dictionary section of the store which is defined by A_{11}, A_{12} and the contents of the counter designated by A_{13}
51	Store to Dictionary	A word in the letter register is stored to the location defined in 50
52	Consult WD	The contents of the letter register are compared with the contents of the word dictionary. If coincidence is obtained, next instruction is taken from A_2 and the address of the word in the dictionary is left in a particular counter called the dictionary counter. If not, next instruction is taken from A_1
53	Consult ID	The comparison is with the idiom dictionary. The other actions are the same as in 52
54	Consult SD	The comparison is with the syntax dictionary. The other actions are the same as in 52
55	Consult X	Not used

Table 15.1
(continued)

Code	Abbreviation	Operation
56	Consult Y	Not used
60	Raise Counter	The content of the counter designated by A_{13} is increased by one
61	Lower Counter	The same counter is reduced by one
62	LR Shift	The contents of the letter register are shifted by A_{13} character; to the left when A_{11} is even, and to the right when A_{11} is odd
63	LR 2 to LR 1	The two letters in the letter register 2 designated by A_{11} are brought to the most significant digits of the letter register 2
64	Acc Minus Jump	The next instruction is taken from A_1 if the content of the accumulator is negative. Otherwise, from A_2
65	Acc Zero Jump	The next instruction is taken from A_1 if the content of the accumulator is zero. Otherwise, from A_2
70	Type out LR	Type the contents of the letter register
71	Type Special Char.	Type the character designated by A_1
72	Clear Counter	The counter designated by A_{13} is cleared
73	Stop	The machine is stopped

5 Conclusion

As the program test is not finished, it is difficult to draw any concrete conclusions on the English-Japanese machine translation. However, the experience of programming shows that there are no essentially difficult problems, except when a word has multiple meanings in the same part of speech. In comparison with the translation between European languages, two problems predominate: the word order difficulties and the problem of auxiliary words peculiar to Japanese, which have no corresponding part of speech in English.

As for the computer Yamato, it will be necessary to increase the storage capacity in the near future. It is being planned to add a photographic permanent store such as is used in the machine of the University of Washington [1] for dictionaries and tables.

The Japanese language, which has been cultivated in an island-country for many years, was largely modified by the introduction of Chinese letters about ten centuries ago. It is felt that there are so many irregularities that the language, and also the letters, have to be modified to ease machine translation. A similar request may arise also for English. It is desirable for the mutual understanding of all of the peoples on the globe that articles in various languages should be written according to rules which are convenient for machine translation.

References

[1] Wall, R. E.: University of Washington Eng. Exp. Station Rep., No. 108, 1956.

[2] Wada, H. et al.: *An electronic reading machine*, these proceedings IV B.

[3] Nishino, H. et al.: *ETL Mark IV, a transistor digital automatic computer*, Journal of the Institute of Electrical Communication Engineers of Japan 1959, to be published.

This page intentionally left blank

II
THEORETICAL AND METHODOLOGICAL ISSUES

This page intentionally left blank

INTRODUCTION

Yorick Wilks

What are the methodological and theoretical issues that are confronted in the selection of papers in this section? We must bear in mind as we answer that question that our decision to group historical and systems papers separately means, inevitably, that some of those issues are also discussed in papers in two other sections. Let us begin an answer by surveying the current MT scene for a moment.

At the time of writing, MT seems to be enjoying a revival in North America (driven like so much else at the moment by the World Wide Web), to be somewhat depressed in Japan compared with five years ago, and to be rather stagnant in Western Europe. Ways forward from here seem to be a better attempt at matching the existing technology to market needs, which are real and expanding, and improving the basic technology by getting above the 65–70% level that customers—of the best systems on a good day—now accept.

What does the last possibility mean in concrete terms and how does it relate to the arguments about principle that have raged through MT's past? Everyone, from every research school, can agree with this position, whether one advocates knowledge-based methods, linguistic purity, statistical methods, or only vastly improving editing tools and interfaces for users. There are now much better decomposable modules for linguistic analysis available: part of speech taggers, automatically and semi-automatically derived lexicons and concordances, effective grammars and parsers far closer to corpora than before. Yet their effect is not apparent in systems on the market. Researchers who help build marketable systems still throw away all their theoretical beliefs and their successes when going to market, as if they themselves do not believe in the effectiveness of their own published work.

Martin Kay once argued that, even if all the problems of syntax, morphology, and computational semantics had been individually solved, it might not improve MT. I am not certain what he meant by that but it is an important idea; if one asks what he thought might still be missing one might list research in:

- the gist, meaning or information content of what a text was about,
- a high enough level theory of the rhetorical structure of texts,
- a computationally accessible representation of common-sense knowledge down to excruciating levels of detail,
- a model of generation that gives priority not to rules but always to the smoothest collocational choices,
- a model of the vagueness and indefiniteness of the boundary of much of the categorical knowledge underlying language and thought.

Some, a few, researchers have attempted to tackle these notions directly over the years, but no one can honestly say they have yet proved that their theories, when fully operational, will lift the magic MT success figures much.

Let us stay at this high level of abstraction and ask if the classic theoretical disputes in the history of MT have been solved or at least put to rest, and if their solution has

affected the current situation much. I suggest that the following is a fairly simple first list of the crucial issues of the last twenty years, most of them illustrated by works reprinted in this part:

1. Representations for MT: the role of knowledge. Bar Hillel argued this was essential for MT but unachievable with computers. AI researchers in MT accepted the premise but believed the task *was* achievable, yet most proprietary MT systems still do not attempt to encode anything resembling a knowledge component beyond their technical vocabulary. It is unclear that this issue, so clear when stated, has yet been settled because it has had little impact on working commercial systems. (See chapter 14, "Mechanical Pidgin Translation" in part I, chapter 23, "Treatment of Meaning in MT Systems" in part II, and chapter 36, "Machine Translation Without a Source Text" in part III.)

2. Representations for MT: the argument for no representation at any level. Data-driven approaches to MT—at least in extreme forms like IBM's statistical model of MT, CANDIDE—deny that any separate representation at all is required: not at the knowledge level nor at any linguistic level. That approach seems to have retreated for the moment in its extreme forms while remaining the principal opposition to representational orthodoxy, even though with little influence on commercial systems. (See chapter 4, "Stochastic Methods of Mechanical Translation" in part I, and chapter 31, "A Framework of a Mechanical Translation Between Japanese and English by Analogy Principle" in part III.)

3. Representations for MT: language-specific or not. This is the interlingual argument versus transfer and direct methods—although no one now seems to defend direct coupling methods much since SYSTRAN declared itself to be a transfer system some years ago. There is little support for true interlinguas (which by definition exclude language specific resources like bilingual lexicons) and the argument is really about interlinguas+bilingual lexicons, since renouncing the use of a bilingual lexicon to restrict generation choices seems hobbling oneself unnecessarily. After that, interlingualism may be a matter of degree, shading into controlled language inputs used as pivot languages and so on. (See chapter 14, "Mechanical Pidgin Translation," in part I; chapter 22, "Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation," and chapter 25, "The Place of Heuristics in the Fulcrum Approach to Machine Translation" in part II; and chapter 34, "The Stanford Machine Translation Project," in part III.)

4. Representations for MT: "high-level non-AI theory." Does MT need "the best linguistics"? This is a hard one as there is no agreement as to what the best linguistic theory is at any given moment. Certainly there has been far more contact in the last ten years between front-line linguistic systems (such as the acronymic FUG, LFG, HPSG, and GPSG type grammars, Berwick principles, even) than in the decades before, which contained only IBM's transformation-based system, a Dutch research system based on Montague semantics and, as always, EUROTRA. I suspect one cannot identify any linguistic theory in many MT systems under serious evaluation or on sale at the moment.

5. MT as a directional task: analysis- or generation-driven. There has been a substantial shift here in the last decade towards the role of generation as an "intelligent" task and one preeminent in the practice of actual translators. Generation seems to have been the most successful part of the data-driven statistical systems. There seems general agreement on this in the research area, though again it is hard to locate this fact within commercial MT systems except perhaps in the few example-driven systems in operation, since these could be said to be driven entirely by past generation

activities. (See chapter 24, ''Where Am I Coming From: The Reversibility of Analysis and Generation in Natural Language Processing,'' in part II.)

6. The scope of a language: the sublanguage issue.

There is some realization now that this has become a non-issue since all MT systems define the scope of what they cover and hence a sublanguage. (See chapter 16, ''Automatic Translation and the Concept of Sublanguage,'' in part II and chapter 36, ''Machine Translation Without a Source Text,'' in part III.)

7. The scope of a language: must dialogue be treated quite differently in MT from running prose? I suspect this is an important distinction and masked by the fact that prose MT and speech-to-speech MT now seem to develop in quite different environments—this hides the fact that the role of pragmatics and the nature of the parsers required for them are probably utterly different. The lack of a production speech-to-speech system, though there are many prose systems, is not due to the weakness of speech front ends alone. (See chapter 20, ''Dialogue Translation vs. Text Translation: an Interpretation Based Approach,'' in part II.)

8. MAT as a distinct and inevitable form of MT. If one accepts Bar-Hillel's argument on impossibility then MAT is really all there is. Computers can certainly provide editing tools and when Kay proposed MAT as the survivable form of MT he probably did believe Bar-Hillel and underestimated the survival potential of low-quality full MT. However, MAT has (with the PC/MS revolution) proved a productive gradualist path towards MT: indeed, in a sense much MT is now *for* translators, and they are an identifiable market. (See chapter 17, ''The Proper Place of Men and Machines in Language Translation,'' in part II.)

9. The problem of extensible or metaphoric lexical meaning. This has been a major independent area of lexical research, which has received a new boost from the clear limitations built into the ''machine-readable dictionary'' movement, which attempted to extract lexicons for MT from existing dictionary sources and did to some extent succeed. But that, in a sense, only highlights how much word sense coverage is not present in existing dictionaries and is not simply the product of an inadequate terminology coverage: i.e., is the product of extension of sense specific to certain domains, text types, etc.? There has been substantial research into whether there are lexical rules for regular, non-domain-specific, extensions of sense or whether there are patterns of metaphor that endlessly repeat, as Lakoff seems to hold, but again there has been no input to systems now available.

In summary, one might say that there is real progress in MT nevertheless, and this list of MT issues does not stay unchanged over long periods of R&D. What is less clear is the ways in which, and the time scale by which, high-quality research reaches products; it is far more mysterious in MT than in most areas of computer-based R&D, such as VLSI or even closer areas like information retrieval and extraction where there seems a reasonably well understood route from tested research to fielded systems.

This page intentionally left blank

16
Automatic Translation and the Concept of Sublanguage

John Lehrberger

1 Introduction

It is common to speak of the language of biophysics, the language of pharmacology, etc. as though there were certain well defined languages used by specialists in various fields. But a glance at technical or scientific writing reveals that the language used is basically a language such as English or French. Even a layman can recognize the language although the presence of special terminology and mathematical formulas may prevent him from understanding the subject matter. If we can recognize that a text is "in English" and yet feel that it is distinct enough to be described as being "in the language of X" (physics, aeronautics, electronics, etc.) then we may be justified in saying that the language of X is a "Sublanguage" of English. In fact, the term *sublanguage* is now used by many linguists investigating texts in specialized fields. And it is within the domain of sublanguages that automatic translation appears to be practicable. A system for translating weather reports from English to French is already in use in Canada (Taum-meteo [1]) and a system for translating aviation maintenance manuals is under development at the Université de Montréal [4]. This paper will examine the notion of Sublanguage, its role within the "whole" language, and its importance in the development of automatic translation.

2 Description of a Particular Sublanguage

2.1 The Corpus

Researchers at TAUM (Traduction Automatique Université de Montréal) have made a detailed study of the properties of texts consisting of instructions for aircraft maintenance. The study was based on a corpus of 70,000 words of running text in English. There were 3,548 different words in the analysis dictionary distributed among the various categories as follows:[1]

(1)	nouns	1714	prepositions	134
	verbs	667	coordinate conjunctions	13
	adjectives	664	subordinate conjunctions	29
	adverbs	168	pronouns	35
	quantifiers	46	articles	15
	numerals	63		

Only base forms are listed in the dictionary (e.g., *adjust* is included, but not *adjusted* or *adjusting*). There are 571 idioms included in these figures, 443 of which are "technical" idioms specific to the subject matter. Examples of these idioms and a discussion of the criteria for listing an expression as an idiom will be given in section 2.5. Further study is expected to result in a reduction in the number of idioms at a later stage.

The categories in (1) are traditional; words (and idioms) are assigned to these categories on the basis of their use in the corpus. E.g., "cool-skin," "gear-driven," "loadcarrying," "following" are all listed as adjectives since they occur only as modifiers of nouns. Further subcategorization, essential for parsing, is obtained by associating syntactic and semantic features with the words in the analysis dictionary. Thus with each dictionary entry there is associated a category, features and complementation (a detailed description of the format of the analysis dictionary is given in [4]).

2.2 Restrictions

2.2.1 Lexical Restrictions Although the corpus contains only 4,876 different lexical items it is estimated that in the set of texts which the corpus represents there may be something like 40,000. Comparing this number with the number of entries in Webster's Third (about 450,000), it is obvious that the vocabulary of this Sublanguage is highly restricted. One needs to describe the parts of the aircraft, the maintenance of hydraulic systems, electrical systems, turbines, etc., and the tools and test equipment required for such maintenance. Certain words are characteristic of this subject matter: aileron, motor, compressor,

jack, filter, check, axial, quick-disconnect. Other words do not occur at all: parsley, meson, seduce, endocrine, hope, think, believe. None of the personal pronouns *I*, *me*, *we*, *us*, *he*, *she* are used here. The sets of words which characterize different sublanguages are not mutually exclusive, however. "Filter," which occurs frequently in this corpus, is also typical of the language of pharmacology. It is not the vocabulary alone which determines a Sublanguage, as we shall see in the following sections, although it is certainly an important factor.

Vocabulary restrictions do not apply to the same extent in all categories. The categories noun, verb, adjective and adverb are most limited while nearly all members of the remaining categories may be found in most sublanguages. E.g., all articles and coordinate conjunctions occur in this corpus. About 70% of one-word subordinate conjunctions occur (we do not expect to find "whilst" or "whereupon") and about 80% of one-word prepositions (nor do we expect "apropos" or "notwithstanding"). This result conforms to the ubiquitousness of "grammatical" words and the fact that the main semantic burden is borne by nouns and verbs. On the other hand there are sublanguages which are characterized by the use of certain archaic or formal grammatical words ("whereupon the Lord commanded") as well as typical nouns and verbs.

2.2.2 Syntactic Restrictions Since the sentences of this corpus are used either to describe the aircraft and related equipment or to give instructions for their maintenance, direct questions do not occur at all (*Do you have your tool kit? *Is the motor turned off?). And tag questions indicate an attitude toward the user of the manual which is unacceptable (*Check the batteries, won't you? *The switch should not be on, should it?), hence they do not occur.[2]

There is no use of the simple past tense in the corpus (*The engine stopped. *High temperatures caused buckling.)

There are no exclamatory sentences (*How powerful the engine is! *What a complex hydraulic system this plane has!)

Other sentence types show the full range of syntactic structures in the corpus: passives, restrictive and non-restrictive relative clauses, extraposition, nominalizations of various types, etc. Long and complicated sentences are common in spite of the "telegraphic style" which characterizes most of the text and the internal structure of the noun phrase is often quite complex:

(2) "This unit contains the fuel metering section, shutoff valve, and a mechanical governor that functions as either an over speed governor for the high pressure rotor or provides manual control when the electronic computer section of the fuel control system is deactivated."

(3) "... a lightweight, two-spool geared transonic-stage, front-fan, jet propulsion engine."

One of the most difficult problems for automatic parsing involves conjunction, with its associated reductions and ambiguities. E.g.,

(4) "Disconnect pressure and return lines from pump." (ambiguous)

Another very difficult problem is the proper bracketing of long sequences of nouns:

(5) "The stability augmentor pitch axis actuator housing support" (see 2.4.3).

The corpus is generously endowed with such features so that parsing is by no means simple in spite of the restrictions mentioned above.

2.2.3 Semantic Restrictions

2.2.3.1 Categorization and Subcategorization We have seen in 2.2.1 and 2.2.2 that the restricted subject matter and the attitudinal relation between text and reader limit the vocabulary and the inventory of syntactic structures in the Sublanguage. But more important than the limitation in size of vocabulary is the reduction in polysemy. In some cases this results in a word occurring in only one category in the sublanguage whereas it may occur in several categories in the Language as a whole. E.g., in this corpus the words in (6) occur only in the categories indicated in parentheses.

(6)	case (N)	**Case* the joint.
	lug (N)	*They *lugged* the equipment from the plane.
	cake (V)	*The pilot likes banana *cake*.
	jerky (ADJ)	*Carry a pound of *jerky* on long flights.
	just (ADV)	*This is a *just* test procedure.
	fine (ADJ)	**Fine* them for smoking. *There is *a fine* for smoking.
	cable (N)	**Cable* the forward compartment.

In other cases the range of meanings of a word within a given category is restricted:

(7)	eccentric (ADJ)	Cannot apply to animate objects (*an *eccentric pilot*)
	ball (N)	Can only be a spherical physical object (*the annual *ball*)
	check (N)	Abstract only (*Cash this *check*.)
	bore (V)	Cannot take human object (*Inaction may *bore* the crew.)
	bore (N)	Cylindrical hole or inside diameter of cylinder (*The pilot is a *bore*)

Since the parser explores the possibility of assigning a structure to a given string of words for each category in which the words occur, a reduction in the number of categories to which the individual words belong results in fewer combinations and less ambiguity. E.g.,

(8)	Check	pump	case	drain	fitting.
	N	N	N	N	N
	V	V	V	V	V

In general English each word of (8) can occur in either category N or V, resulting in thirty-two paths to be explored. Of course all but one of these should be rejected by the parser (the combination in which "check" is a verb and the remaining words are all nouns). But since "case" is not used as a verb in the corpus it is listed in the analysis dictionary only as a noun. This alone reduces the total number of combinations to be tested in (8) from thirty-two to sixteen.

Consider the ambiguity in (9):

(9)	Case	ejection	door	locks	immediately.
	N	N	N	N	ADV
	V			V	

In general, either "case" or "locks" may be taken as the verb. However, in this corpus "case" occurs only as a noun. "Locks" is therefore the only candidate for a verb and "case ejection door" is the subject noun phrase. The parser is then relieved of the responsibility for deciding that maintenance personnel are not instructed to case the ejection door locks.

Restriction of the semantic range of a lexical item, even when it does not reduce the number of categories to which the item is assigned, is extremely useful in parsing. E.g., in (10) "cooling" may be taken either as a modifier of "purposes" or as the gerundive form of the verb "cool" whose object is "purposes."

(10) (A small heat exchanger) uses engine fuel for cooling purposes.

It will be obvious to the reader that "purposes" is not the object of "cooling," but how does the parsing machine know it? In these texts only concrete things are cooled (not tempers, etc.), hence we need only specify in the dictionary entry for the verb "cool" that its direct object must have the feature CONCRETE. If we were designing a parser for all English this would not suffice. The subcategorization required to establish all necessary cooccurrence restrictions for the whole language would be very fine indeed. Even in a Sublanguage the elimination of ambiguities is a serious problem.

2.2.3.2 Specificity These texts are characterized by the absence of generic reference of the form "the+N". In the language of biology we have "*The dolphin* is a mammal." In a history text we may find "The invention of *the wheel* was a crucial step." But in these aircraft maintenance manuals the sequence "the+N" is specific. E.g.,

(11) *The oil tank* is not a component of *the engine*.

(12) *The computer* provides increased fuel scheduling.

"The wing," "the radio," "the engine," "the wheel," etc. are all specific references. The manual differs from a textbook which may be concerned with theoretical concepts and general definitions. Whereas a textbook on motors and generators may contain a statement like (13):

(13) The motor is a machine that converts electrical into mechanical energy.

an aircraft maintenance manual contains statements like (14):

(14) The motor is a constant-displacement piston type.

Thus there is no ambiguity in this corpus involving generic versus specific reference. A further consequence of this fact is illustrated by the sentence

(15) Clean reservoir system.

We may assume deletion of the definite article has taken place if we wish to compare (15) with the corresponding sentence in "standard English":

(16) Clean the reservoir system.

Instructions for maintenance and repair must be specific; one does not expect to find "Clean a reservoir system." Of course, we do not really have to recover deleted articles to *understand* sentences such as (15). We merely need to recognize the general principle concerning a specific reference and then accept the fact that (15) is a normal acceptable sentence

in this sublanguage (see section 2.3.1). (However, the French translation does require a definite article, hence we must recover it for the purpose of automatic translation.)

2.2.3.3 Semantic Features The semantic restrictions imposed by the subject matter are reflected in both the number and kinds of semantic features needed for parsing. Many nouns which designate either concrete or abstract objects in the language as a whole are used only concretely in this Sublanguage; e.g.,

(17) air, battery, dirt, machine, flap, flash, post, rod, solution, speed, spring, tool, net, web, race.

The same is true of words that may be used for either human or non-human objects. None of the following words which appear in the corpus designate human beings or parts thereof:

(18) agent, body, boss, buffer, crank, elbow, governor, joint, nut, page, selector, starter

Verbs are likewise restricted in the kinds of subjects and objects they can take:

(19)

charge	object [+CONCRETE]
circulate	subject [+FLUID] (intransitive)
divert	object [+FLUID]
function	subject [+PART] (i.e., part of the aircraft or related equipment)
top	object [+CONCRETE]
die	subject [−ANIMATE]

The features MALE, FEMALE are not relevant in the corpus. The feature HUMAN has been used on only a few nouns in the parser although many verbs are marked as taking HUMAN subjects since this is implied by the use of imperatives throughout the text:

(20) Check fan blade clearance.
Adjust pump pressure control valve.
Remove and discard gasket.

Thus the feature HUMAN is used mainly in signaling implied subjects rather than in testing nouns in the text as possible subjects of nearby verbs.

The degree of semantic restriction in the Sublanguage has a bearing on the manner of representing semantic features. In fact, two types of representation have been considered for the parser, which we may call unary (*F) and binary (+F or −F). The criterion for admitting a noun having the set of unary features $\{{}^*F_1, {}^*F_2, \ldots, {}^*F_n\}$ as the kth argument of a verb whose kth argument position is assigned the set of unary features $\{{}^*G_1, {}^*G_2, \ldots, {}^*G_n\}$ is that the two sets have a non null intersection. This means that if the nth argument of a verb can be either *CONCRETE or *ABSTRACT then both of these features must be listed in the nth argument position of the verb in the dictionary. And if a noun may be either *CONCRETE or *ABSTRACT then both features must be entered with that noun in the dictionary. This would seem to result in a great deal of redundancy since there are so many nouns which may designate either concrete or abstract objects and so many verbs whose arguments may be either *CONCRETE or *ABSTRACT. The same is true for many other features as well.

The alternative is to use binary features along with the following conventions:

(21)
(i) A noun is marked +F if it *always* has the feature F and −F if it *never* has that feature; otherwise it is not marked for F in the dictionary.
(ii) If the nth argument position of a verb is marked αF it can only take arguments marked αF, where α is either + or −; the nth argument position of the verb is not marked for F if it can take either +F or −F arguments.
(iii) A noun is admitted as nth argument of a verb provided there is no feature F such that the nth argument position of the verb is marked aF and the noun is marked $-\alpha$F.

At first sight it appears that such use of binary features would result in overall economy. However, the semantic restrictions in this Sublanguage result in many nouns being marked only +F (or −F) and many verb argument positions being marked only +F (or −F), as was illustrated in (17)–(19). Consequently the advantage of not having to mark a large number of nouns and verb argument positions for certain features is lost. At the same time, any feature which is rarely used will have to be entered on all those nouns and verbs where it is not relevant. E.g., the small set of nouns including *air*, *oil*, *water*, etc. would be marked +FLUID and all other nouns in the dictionary would have to be marked −FLUID as well as all verb argument positions which do not accept +FLUID arguments. However, if unary features were used then *air*, *oil*, *water*, etc. would be marked *FLUID and the majority of nouns would not be

marked at all for this feature. Likewise for verb argument positions.

Presently the unary method of representing features is used in the parser. At any rate, it is interesting to note the effect of semantic restrictions within a Sublanguage on the kind of semantic representation which is used.

2.3 Reductions

2.3.1 Omission of Definite Article One of the most frequent reductions found in this corpus is the omission of the definite article:

(22) Check indicator rod extension.
One system provides air for bearing compartment sealing.

But such reduction does not always take place, as can be seen in the following sentences which are also found in the corpus.

(22′) Check *the* ground test system.
Check *the* control stick breakout.
All controls for *the* air conditioning system are located in *the* front cockpit.
Separate outlets are provided for *the* engine and handpump.

It does not seem to be the case that in some contexts the definite article is always omitted while in others it is not. We can only say that it may be omitted and very often is. Yet, in spite of this, *the* is the most frequently occurring word in the corpus (2,925 occurrences). No definitive study has been made of the environments where its omission is most likely to take place. From the point of view of the parser, allowance is made for the fact that it may not be present where it is expected in standard English, but no attempt is made to predict its presence or absence.

2.3.2 Omission of Copula In standard English the copula BE may or may not be used in certain contexts:

(23) (i) The book (which is) on the desk.
(ii) We considered it (*to be*) unreliable.

We may question whether the shorter forms are "reductions" or simply paraphrastic alternatives, but in the corpus there is another type of construction which seems clearly to be a reduction involving omission of BE:

(24) (i) Check reservoir full. (Check that the reservoir is full.)
(ii) Check fluid level above REFILL mark. (Check that the fluid level is above REFILL mark.)

There is a class of verbs (believe, consider, find, etc.) which can take a noun phrase +*to be* ... as complement (as in 23ii). When *to be* is not present the complement may consist of a noun phrase followed by an adjective phrase. But *check* does not belong to this class of verbs, hence the construction in (24i) is peculiar to these texts. Both (24i) and (24ii) occur frequently in the corpus. However, there are similar sentences which do contain the copula:

(25) (i) Check that fuel systems are full.
(ii) Check fluid level indicator is registering correctly.
(iii) Check that fuel pressure is between 45 and 55 PSI.

As with the definite article we see that the omission of *be* is not obligatory in sentences like (24). It does happen often enough to be considered characteristic of these texts, but it is optional in those contexts where it occurs. The copula is also omitted from progressive forms, as in (26):

(26) Pump not delivering fluid.

2.3.3 Omission of That Complementizer A comparison of (25ii) with (25i) and (25iii) shows that the omission of *that* as a sentence nominalizer is optional. This is common in standard English with verbs like *suppose*, *know*, *hope* (I suppose/know/hope the fluid level indicator is registering correctly), but not with the verb *check* (*we are checking the indicator is working).

2.4 Frequently Occurring Forms

2.4.1 Imperative Imperative sentences abound in the corpus. This is to be expected since a maintenance manual, like a cook book, is primarily concerned with instructing the user in the performance of certain actions (Check ..., Adjust ..., Turn ..., Remove ..., Insert ..., etc.). Were it not for the fact that these manuals also describe parts of the plane and how these parts function, nearly every sentence would be in the imperative. The significance of the imperative in characterizing this Sublanguage is not simply that it occurs, but that it occurs so often.

2.4.2 Non-Predicative Adjectives There are many adjectives in the corpus which never occur in predicate position. They are marked with a feature ATRIB

in the parsing dictionary and constitute 25% of all adjective entries. Some examples are given in (27).

(27)

A.	B.
actual	nickel-cadmium
chief	piston-type
consequent	pressure-regulating
entire	anti-stall
respective	single-point
	non-priority

Those listed in column (B) are particularly important in characterizing this Sublanguage. They deal specifically with the subject matter of aircraft maintenance whereas those in column (A) are of a more general nature. There are a number of productive types involved in (B):

(28) X-type
X-Ving
anti-X
X_{num}-Y
non-X

Presently all the adjectives in (27) are listed in the parsing dictionary. However, since those in (B) are productive types it might be preferable to separate the components at pre-edition (see [4]) and analyze the resulting string in the parser. This matter is under study.

In addition to being non-predicative these adjectives are not inflected for comparative or superlative (*chief*er*, *pressure-regulating*est*). We might question whether they should be considered as forming a subcategory of the adjective class or as simply a separate class of prenominal modifiers in this Sublanguage (see 4.2).

The corpus contains many compounds consisting of a numerical expression followed by either a measure unit (29A) or a certain kind of noun which may be considered a measure unit in the proper context (29B):

(29)

A.	B.
115/200-volt	3-phase
0.0045-inch	19-cell
10-micron	2-stage
1000-hour	eighth-stage
15-ounce	two-lobe
11-ampere-hour	three-way
110-to-infinite HERTZ	two-spool
	three-axis

The nouns following the dashes in (B) are not, strictly speaking, measure units; but in this Sublanguage they are used as such: *phase* is a measure unit with respect to *generator*, *cell* with respect to *battery*, *lobe* with respect to *cam*, etc. All compounds of this type should be separated into their components at pre-edition, analyzed by the parser, and assigned a feature MP (measure phrase). They should not be entered as individual lexical items in the dictionary since the numerical portion is of arbitrary size. Just as numerical expressions in general must be parsed, so must these measure compounds.

One of the conventions followed in these texts is to place a hyphen between a numerical expression and following measure unit when the compound is used as a prenominal modifier, and to write the measure unit in the singular; otherwise there is no hyphen and the measure unit may be pluralized (a three-stage turbine, the turbine has three stages). The hyphenated measure compounds behave like the non-predicative adjectives described above, hence they should be treated as adjective phrases by the parser and so labeled. The numerical component might suggest treating them as quantifiers but unlike most quantifiers they occur in the characteristic adjective position between article and noun and do not require pluralization of a following count noun in this context.

2.4.3 Noun Sequences A major feature of the corpus is the presence of many long strings of nouns, or nouns and adjectives, within nominal groups:

(30)
(i) external hydraulic power ground test quick-disconnect fittings
(ii) fuselage aft section flight control and utility hydraulic system filter elements
(iii) fan nozzle discharge static pressure water manometer
(iv) landing gear, flight controls, speed brakes, engine air bypass flaps, and nose steering systems
(v) stabilizer power control No. 2 system return line check valve failure

This phenomenon is a result of the need to give highly descriptive names to parts of the aircraft in terms of their function in the aircraft and their relation to other parts. It is likely to occur in any texts describing very complex machinery containing a large number of specialized parts.

The segment of such a noun phrase from the first adjective or noun to the last noun is referred to here as an *empilage*. It does not include initial determiners or quantifiers. In the corpus there are about 4,400 different empilages, many of them occurring numerous times. They present a major problem in parsing the nominal group.

The proper bracketing of an empilage requires an understanding of the semantic/syntactic relations between the components. Thus in (31)

(31) main fuel system drain valve

we must know that *main* applies to *fuel system*, not to *fuel* or *drain valve*, and that (31) refers to a valve whose function is to drain the main fuel system, whereas in (32)

(32) system main drain valve

main applies to *drain valve*, not to *drain*, and (32) refers to the main valve whose function is to drain the system.

The problems involved in parsing empilages are similar to those encountered by linguists attempting to explicate the formation of NOUN+NOUN compounds. Recent investigators have proposed various semantic and syntactic relations between the components as well as "underlying structures" for analyzing these compounds. E.g., Levi [6] proposes a small number of relations (deletable predicates) such as CAUSE (nicotine fit), HAVE (fruit tree), USE (steam engine), FOR (communications system), SUBJECTIVE NOMINALIZATION (cell decomposition), OBJECTIVE NOMINALIZATION (traffic control), etc. in terms of which most NOUN+NOUN compounds may be derived. Downing [2] suggests twelve relations that should be included in any such inventory [WHOLE-PART (duck-foot), COMPARISON (pumpkin bus), TIME (summer dust), PLACE (Eastern Oregon meal), etc.], but she claims that "the semantic relationships that hold between the members of these compounds cannot be characterized in terms of a finite list of 'appropriate compounding relationships'". The study of empilages at TAUM resulted in the definition of about 50 such relations, including some of those mentioned above (HAVE, WHOLE-PART, PLACE, SUBJECT, OBJECT, etc.). This list of semantic/syntactic relations may turn out to be sufficient for analysis of empilages in the Sublanguage under investigation, but no claim is made for the whole language. If Downing is correct, success in finding a finite set (at least small enough to be useful in automatic parsing) may depend on just such limitations as are encountered in a sublanguage.

An example of the constituent structure of an empilage showing the relations between constituents is given in (33) (see figure 16.1).

We have not been able to define a set of relations which are mutually exclusive, as is evidenced in (33) by FOR, O and HAS, O.

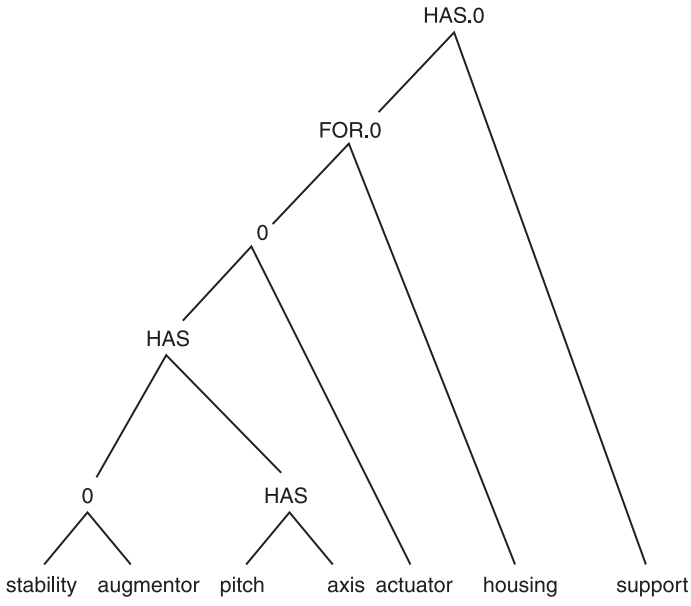


Figure 16.1
Stability is grammatical OBJECT of *augment*; *pitch* HAS an *axis*; *stability augmentor* HAS a *pitch axis*; *stability augmentor pitch axis is* OBJECT of *actuate*; the *housing* is FOR the *stability augmentor pitch axis actuator* (which is also OBJECT of *housing*); the *stability augmentor pitch axis actuator housing* HAS a *support* (and is also OBJECT of *support*).

2.5 Idioms

For the purpose of this discussion we will use *idiom* as a technical term for any multi-word expression which is entered in the dictionary (except for left members of normalization rules such as "for comparison of == to compare"). Several criteria may be used to determine whether a given expression in the Sublanguage is to be entered in the parsing dictionary as an idiom.

(I) The meaning of the expression is not predictable from the meanings of its components: with respect to, nose gear, finger tight.

(II) Translation idioms. The corresponding expression in the target language is not predictable by the usual rules for associating structures in the target language with those in the source language. Aspect ratio (Fr: allongement), DC power (Fr: courant continu), buttock line (Fr: section longitudinale).

(III) The expression occurs so often in the Sublanguage that it "feels like" a compound word: landing gear, filter element, relief valve. From the point of view of automatic translation it may be more economical to list these in the dictionary than to parse them. That depends on parsing strategies and the desirability of limiting the size of the dictionary.

(IV) The expression occurs very rarely and its parsing would require undesirable changes in strategies developed to handle the majority of cases in the Sublanguage. E.g., *right and left of center* occurs only once in the corpus ("right and left of center positions"), while *right* and *left* occur quite often. As things now stand, if the expression is not listed as an idiom the parser will spend a lot of time looking for of-complements of *right* and *left* when these words occur elsewhere.

(I) is the usual criterion for idioms, (II) is relative to translation into another language, (III) and (IV) are relative to strategies for automatic parsing. The creation of idioms for the purpose of automatic parsing will be discussed further in section 3.4.

2.6 Text Structure

2.6.1 Gross Structure Texts in this sublanguage are divided into numbered sections each of which deals with a specific part of the aircraft. The occurrence of a polysemous word in a particular section may signal a specific meaning for that word. E.g., *capacity* refers to volume in the hydraulic system and to farads in the electrical system or electronic equipment. In some cases this affects the translation of the word. Thus *valve* is translated into French as *clapet* when it occurs in the section on hydraulics and as *soupape* or *valve* in the section dealing with motors. Since *valve* is one of the most frequently used nouns in the corpus (716 occurrences, including the plural) it would be quite useful at transfer to have available not only the sentence in which the word occurs, but an indication of the section in which the sentence occurs as well.

2.6.2 Linking Devices Neighboring sentences in the corpus are "linked" in various ways to provide textual cohesion. Repetition of a word or phrase is common:

(34) Install rotor on *shaft*, then align index marks on
(1)
inner races of *bearing*. Position *bearing* on *shaft*
(2) (3) (3) (1)
with vendor identification marking on outer race on same side as puller groove on
inner races.
(2)

When a noun phrase occurs in one sentence it is often shortened in succeeding sentences rather than repeating it in full:

(35) Remove *jumper hose* from *pressure in line*. Cap
(1) (2)
line and *hose*.
(2) (1)

(Note: *in line is* not a prepositional phrase; an in line is a line bringing something in.)

Such reductions complicate the identification of coreferential noun phrases in different sentences even though the head of the NP is retained. And when more than one noun phrase is repeated, as in (34) and (35), the order of occurrence may be different.

Pronouns may also operate across sentence boundaries:

(36) The main system relief valve is located on the left side of the engine compartment, just forward of the hydraulic reservoir. It is adjusted to provide automatic relief when hydraulic pressure in the system exceeds 1225 (+−25 PSI).

It refers to *the main system relief valve*, not to *the left side of the engine compartment* or *the hydraulic reservoir*, although the latter two are closer to *it*. In this case we know that valves, not reservoirs or compartments are adjusted, and since *it* is the object of *adjust* in the second sentence, the referend of *it* in the first sentence must be *valve*. In order to make use

of this information in a parser nouns could be subcategorized on the basis of their naming things which can be adjusted (perhaps by assigning a feature ADJUSTABLE). This alone would not be sufficient since more than one adjustable item may be mentioned in the first sentence. Semantic analysis of a text, even when restricted to a Sublanguage, calls for considerable subcategorization.

Repetition with "adjustment" of grammatical category is also used as a linking device. Nominalization is one of the most common adjustments:

(37) Vent manifold may be leaking. This leakage will allow ...

Sometimes there is implicit reference to elements in the preceding sentence without the use of proforms or repetition (unless one's theoretical framework makes it more convenient to consider this a case of repetition with reduction to zero):

(38) Remove and inspect the fuselage aft section flight control and utility hydraulic system filter elements. If found to be highly contamined, clean and reinstall, then remove and inspect all flight control actuator filter elements. If found to be highly contaminated, clean and reinstall, then remove and inspect all hydraulic system restrictors. If restrictors are found to be highly contaminated, clean and reinstall.

The object of *find*, *clean*, *reinstall* in the second sentence is in the first sentence and the object of *find*, *clean*, *reinstall* in the third sentence is in the second, but the object of these verbs in the fourth sentence is present in the fourth sentence owing to the repetition of *hydraulic system* restrictors (with reduction), retaining only the head, *restrictors*.

There are many lists and tables in the corpus. References to them (or to particular sentences in them) in other parts of the text often results in direct linking between non-contiguous sentences. The internal structure of a list may disambiguate an expression contained in it. E.g., consider (39) and (40):

(39) Correct wiring.

(40) Bleed fittings on brake assembly.

Since (40) begins with a capital letter and ends with a period we might assume it is a sentence instructing the technician to bleed the fittings. However, it occurs in the second column of a list of components and their bleed points which is headed COMPONENT BLEED POINT and each expression on the right is the name of a location, not an imperative sentence. Now we have to be suspicious of (39) which may be simply an adjective+noun combination in spite of the initial capital and final period. On examining the structure of the list in which (40) occurs we find that it contains three columns:

PROBABLE CAUSE	ISOLATION PROCEDURE	REMEDY

The third column, consisting of imperative sentences (Clean ..., Install ..., Remove ...), includes (39) which is therefore an instruction to correct the wiring (V+N).

These examples show that semantic and grammatical analysis of a text (or even a sentence) requires looking beyond the boundary of the individual sentence. A unit of text larger than the sentence seems to be needed. The use of such a unit was considered in the development of the present system at TAUM but was rejected for reasons of economy. This does not preclude use in future development as it is both desirable and possible on theoretical grounds.

2.7 Odds and Ends

2.7.1 Numerical Expressions and Reference The corpus makes much use of numerical expressions, either spelled out (secure with two attaching bolts) or written with Arabic numerals (gauge should read 1000 PSI). There are certain rules governing the representation of numbers in these texts: spell those from zero to nine except for percentages (5%), numbers in compound adjectives (two 3-phase generators), all numbers in a sequence if one of them exceeds 9 (position clamps 8, 11, 21, 24, 30 on harness), etc. However, all numerical expressions are represented by Arabic numerals after parsing in the present translation system used at TAUM since this is more convenient at the transfer stage.

There are many expressions consisting of a mixture of numerals, letters hyphens and slashes, which are called references ("refer to EO 15-70-5A/2"). These have an internal structure which is semantically significant, but for the purpose of translation they keep the same form.

2.7.2 Labels Frequently a word in the corpus refers to a label on a part of the aircraft or related test equipment. These words, indicated by spelling with all capitals, are not to be translated.

(41) Set switch to ON.

(42) Ensure that the PITCH CONT switch is ON.

In (41) ON is simply a label, but in (42) it also serves one of its normal grammatical functions as an intransitive preposition (or prepositional adverb). The use of this kind of ambiguity in these texts reflects the general tendency to be as concise as possible. Of course, since the labels are not to be translated this can be troublesome: the switch is ON = l'interrupteur est sur "ON." The systematic ambiguity does not hold in translation, hence the restructuring in French with the addition of *sur*. It is as though the English sentence had been "the switch is on ON."

2.7.3 N-Ving, N-Ved There are many compound words in the corpus consisting of a noun followed by the present or past participle of a verb (gear-driven, air-separating, cockpit-mounted, motor-operated, seat-adjusting, spring-adjusting, spring-loaded, etc.). The noun usually names a part of the aircraft and the verb describes an operation on or by that part. These compounds are entered in the dictionary as adjectives when there is no corresponding verb. Consider the following example involving *gear-driven:*

(43) A spinner hub and an axial flow fan are gear-driven by the low pressure spool.

Since *by the low pressure spool* is agentive, not locative, (43) appears to be a passive and *gear-driven* the past participle of a verb *gear-drive*. But *gear-drive* does not appear as a verb in these texts (*X gear-drives Y). Hence we accept the structure *N be A by N* where *by N* is agentive.

3 Practicability of Automatic Translation

3.1 Formal Grammars for Natural Languages

Perhaps the problem of designing an automatic translation system for a natural language may be viewed more clearly from the perspective of attempts to write formal grammars for natural languages. It is precisely when we try to formalize our knowledge of a language that the difficulties begin. Generative grammarians in particular have put an enormous amount of effort into the formalization of rules of grammar. Their lack of success so far in producing a set of rules that will generate all and only the sentences of a natural language in its entirety hardly seems encouraging to researchers in automatic translation trying to devise a set of rules that will analyze any sentence in one language and generate the corresponding sentences in another. In fact, the prospect may seem even dimmer when we consider that generative grammarians usually aim only for a description of the "standard language" or the language of an "ideal speaker in an ideal community"; presumably a natural language in its entirety includes arbitrary discourse, much of which lies outside these domains.

Is it then realistic to expect success in automatic translation given the difficulty of writing a formal grammar for even one language? One may reply that automatic translation from L_1 to L_2 does not require complete grammars of L_1 and L_2, only context sensitive transfer rules to obtain the proper lexical items in L_2 and some rules for restructuring the resulting string of lexical items in L_2. Of course, the terms *context sensitive* and *restructuring* in themselves indicate the need to recognize the possible structures in which the lexical items of L_1 and L_2 occur. Experience at TAUM, even with a very limited corpus, has demonstrated that an extremely fine grammatical analysis of both languages is required (especially the source language) in order to translate say 80% of the number of sentences in a text. The system currently in use at TAUM parses the sentences of the source language and puts them in a normalized form indicating their grammatical structure. The "normalized structure" is a tree with labeled nodes and includes semantic as well as syntactic information. Transfer rules map these trees onto other trees containing the proper lexical items of the target language. Rules are then applied which map the trees onto sentences in the target language. Parsing, transfer and generation all require detailed analysis of grammatical structure. The problem of writing rules for a system of automatic translation cannot be separated from the general problem of writing formal grammars for particular languages. The solution in the case of automatic translation seems to lie in restricting one's attention to sublanguages.

3.2 Text Norms

The authors of maintenance manuals, cook books, articles in scientific journals, etc. are generally guided by norms for writing in their particular fields. In some cases guidelines are made explicit. Thus criteria for the texts described in section 2 are given in a booklet titled "Format and Style Guide." These norms do not themselves constitute a grammar—that can only be determined by examining the texts. But they do indicate certain regularities not present throughout the whole language, thus simplifying the task of writing formal grammars for texts in specialized fields.

The existence of norms for texts in certain fields, the reduction in polysemy resulting from semantic restrictions, the limited vocabulary, and the syntac-

tic restrictions generally encountered all combine to make automatic translation practicable for sublanguages. An example of a working system is given in the next section.

3.3 TAUM-METEO

A system for automatic translation of weather reports from English to French is now in use in Canada ([1]). The Sublanguage in this case has a very small vocabulary and is characterized by telegraphic style. Because of the telegraphic style verbs appear only in the present participle or past tense forms. These factors make it more economical to include morphological variants in the dictionary instead of listing only the base forms and performing a morphological analysis.

The syntax is highly restricted: no relative clauses or passives, omission of copula, no use of articles, etc. Consequently syntactic analysis depends very much on semantic subcategorization, as can be seen by the five sentence types recognized in this system.

(44) (i) place names preceding the forecast
RED RIVER
INTERLAKE
(ii) meteorological conditions for the day
MAINLY SUNNY TODAY
WINDS 25 KM PER HOUR
(iii) statement of maxima and minima
HIGHS TODAY 15 TO 18
LOWS TONIGHT NEAR 3
(iv) outlook for next day
OUTLOOK FOR THURSDAY ...
CONTINUING MAINLY SUNNY
(v) heading of bulletin indicating origin
FORECAST FOR MANITOBA ISSUED BY ENVIRONMENT CANADA AT 6 AM CST APRIL 8TH 1976 FOR TODAY AND FRIDAY

(This is a fixed form; only place names, dates and times change.)

METEO is by no means as complex as the system required for the texts described in section 2, but it does demonstrate the feasibility of automatic translation. A complete description of TAUM-METEO is given in [1].

3.4 Idioms

In 2.5 we examined four criteria for entering strings of words in the dictionary instead of submitting them to analysis by the parser. It might be thought that parsing could be greatly simplified by entering many strings in the dictionary even though they do not meet those criteria, especially noun sequences. The dictionary would then be rather large, but by removing much of the burden from the parser where theoretical problems in linguistics are still a major stumbling block the translation of arbitrary texts in a language would seem to be a more reasonable goal. However, it is difficult to imagine just how large a dictionary would be required to eliminate major parsing problems. There seems to be hardly any limit on the number and size of noun sequences possible in a language, judging from the corpus described earlier (2.4.3). Furthermore, a string of words which forms a noun phrase in one context may occur in other contexts where the words have different meanings or belong to different categories and are not even within the same constituent. E.g.,

(45) Locate all check points.

(46) Check points for pitting.

Listing *check points* in the dictionary as an idiom of category N would simplify the parsing of (45) but prevent correct analysis of (46) where *check* is a verb and *points* its object. Furthermore, *points* has a different meaning (as well as a different French translation) in (45) and (46).

Few word sequences are idioms in *all* contexts in a Sublanguage, and even fewer in all contexts in the language as a whole. When a suspected idiom is encountered it is necessary to check that it really is an idiom in that context (see [4] for discussion of treatment of "potential idioms" in TAUM system). An expression which forms an idiom in all contexts in a Sublanguage when its components are contiguous may also occur with the same meaning but with its components separated under certain conditions. This is especially true when conjunction reduction is involved, and it poses a problem in parsing. Suppose, e.g., *in spite of* is entered in the dictionary as an idiom and the parser encounters

(47) He acted without malice in spite and because of her threat.

It is desired to recognize *in spite of* as a unit, but *of* is separated from the rest of the expression. One strategy for putting the pieces back together might be to mark *spite* so that its occurrence immediately after *in* triggers a search for *of* in case a conjunction follows *spite*. More generally, a strategy is needed for reconstituting split idioms.

Problems like these multiply rapidly when a broader range of texts is taken into account and, cor-

respondingly, parsing strategies required to deal with them increase in number and complexity. But when confined to a sublanguage such problems appear to be manageable. E.g., although the idiom *in spite of* could occur in a maintenance manual (the motor may continue operating in spite of fuel leakage), the word *spite* will not occur in the sense of malicious intent and *in spite* will not occur as a prepositional phrase. Thus, having encountered *in spite*, *of* is sure to follow, immediately or otherwise. There can be no ambiguity involving *in spite* and the parser may proceed with confidence to rejoin the components if they are separated.

Experience at TAUM indicates that even in a very restricted sublanguage potential and split idioms constitute a hazard in parsing. Clearly, the creation of idioms is no cure-all in automatic translation.

3.5 Recognition and Generation

A grammar for a natural language may be constructed for the purpose of generating all and only the acceptable sentences of the language or for the purpose of "recognizing" a given string as a sentence and assigning a structure to it. Grammars of the latter type, which we may call recognition grammars, are used for parsing. Normally the input to the parser in a system for automatic translation consists not of arbitrary strings whose sentencehood must be determined, but acceptable sentences whose structures are to be determined. It would be nice to have a machine which could decide for an arbitrary string of words whether or not it is a sentence and assign it a structure, but this is not necessary. In order to parse sentences already assumed to be grammatical one needs strategies for locating verbs and their complements, assigning words to various categories depending on context, assigning constituent structure, etc. This goal seems to be within reach in the domain of sublanguages.

Just as parsing begins with sentences of the source language, so generation of sentences in the target language begins with fully analyzed sentences, i.e., the output of the parser. Words have been assigned to categories, constituents determined, semantic features inserted, etc. Lexical items of the source language must now be replaced with those of the target language and many structural changes effected in the process of generating sentences in the target language. But, difficult as this may be, it is by no means as difficult as starting from the semantic representations, deep structures, or other abstract objects currently employed in many generative grammars and generating all and only the sentences of a language. If the source sentences can be parsed, it's a fair bet that the corresponding target sentences can be generated.

4 The Concept of Sublanguage

4.1 Characteristics

It should be clear from the preceding discussion that a Sublanguage is not simply an arbitrary subset of the set of sentences of a language. Factors which help to characterize a Sublanguage include (i) limited subject matter, (ii) lexical, syntactic and semantic restrictions, (iii) "deviant" rules of grammar, (iv) high frequency of certain constructions, (v) text structure, (vi) use of special symbols.

(iii) refers to rules describing sentences which, though quite normal in a given Sublanguage, are considered ungrammatical in the standard language. Such sentences must be considered grammatical in the sublanguage. (iii) also refers to rules describing cooccurrence restrictions within a Sublanguage that do not exist in the standard language. E.g., in the Sublanguage described in section 2 there is a subclass of adjectives that do not occur with animate nouns (e.g. eccentric pilot). The rule which in the Sublanguage states that "eccentric pilot" is not permitted does not exist in the standard language. It follows that a Sublanguage grammar is not a subgrammar of the standard language. Z. Harris states the matter somewhat differently in [3], p. 154: "... sublanguages can exist whose grammar contains additional rules not satisfied by the language as a whole". (My reason for using "standard language" rather than "language as a whole" appears in 4.3.) Harris claims that sublanguages are closed under transformations (p. 152): "Certain proper subsets of the sentences of a Sublanguage may be closed under some or all of the operations defined in the language, and thus constitute a Sublanguage of it." This notion of Sublanguage is like that of *subsystem* in mathematics. For example, given an algebra $\langle A, f_1, \ldots, f_n \rangle$ where A is a set closed under the operations $f_1, \ldots, f_n$, then a subset of A closed under the same operations forms a subalgebra of $\langle A, f_1, \ldots, f_n \rangle$.

4.2 Cooccurrence and Subcategorization

If a Sublanguage has a grammar of its own which is not just a subset of the rules of grammar of the standard language, it follows that the categories and subcategories of the standard language may not suffice for a grammar of the Sublanguage. This is particularly true of the subcategories needed to state cooccurrence restrictions.

In the work on noun sequences at TAUM relations between nouns were defined on the basis of their behavior in the Sublanguage concerned (2.4.3). (Actually adjectives are included, but the discussion here will be limited to nouns.) Each such relation *R* defines two subsets of nouns, namely the domain and range of *R*. E.g., the relation *F* (*xFy* iff *x* is the function of *y*) has in its domain *access*, *balance*, *check*, *filter*, *installation*, *pickup*, *reduction*, *safety*, etc., and in its range *aircraft*, *bar*, *compound*, *fixture*, *installation*, *lug*, *pipe*, *runs*, etc. The two sets need not be mutually exclusive, as *installation* shows (*installation kit*, *control installation*).

From another point of view, each noun has a left-hand relation-set (the set of all relations having the noun in their range) and a right-hand relation-set (the set of all relations having the noun in their domain). Thus *kit* has *F* in its left-hand relation-set, *control* has *F* in its right-hand relation-set, and *installation* has *F* in both relation-sets.

In order to obtain the correct bracketing of a sequence of nouns it is essential to know the relations that each noun in the sequence can bear to other nouns in the texts under consideration. Now suppose a given noun *n* can bear a certain relation *R* to an immediately following noun. This does not mean that *n* bears that relation to *any* noun that happens to occur immediately following it. For example, although *installation* indicates function in *installation kit* and *installation procedure*, it does not in *installation difficulty* (installation is not the function of difficulty). Thus the subclass of nouns to which *installation* can bear the relation *F* (in the Sublanguage) must be specified, and this is also true for other words in the domain of *F* (*access*, *balance*, *check*, etc.). One way to make such information available to the parser is to indicate in the dictionary entry of a noun all the relations of this type in which the noun participates in the Sublanguage as well as the appropriate subclasses in each case. This may not be an unreasonable task if the number of relations required for the sublanguage is not too great. Of course, noun entries in the dictionary do become fairly complicated and nouns then have a "complementation" similar to that of verbs. The entry for *installation* would specify that the noun can be either abstract or concrete, that when it is abstract it can bear the relation *F* to any member of a certain subclass of nouns occurring on its right (as in *installation kit*, *installation procedure*, etc.), that it can bear the grammatical relation OBJECT to any noun of a certain subclass occurring on its left (as in *pump installation*, *filter installation*, etc.), and that it has certain additional properties when it is concrete rather than abstract.

The whole question of assigning such noun complementation in the dictionary to indicate possible semantic/syntactic relations between nouns (and also noun-like adjectives) is now under study. Clearly, the implementation of such a system depends on a fine subcategorization of the class of nouns, and this subcategorization must be based on a careful study of cooccurrences within noun sequences in the Sublanguage concerned. Although the relations in terms of which these subclasses are defined are of a general nature (FUNCTION, PART-OF, SUBJECT, OBJECT, etc.), the subclasses themselves are specific to the Sublanguage.

4.3 Sublanguages and the Language as a Whole

It is not known how many sublanguages exist in a given language. They are not determined a priori but emerge gradually through the use of a language in various fields by specialists in those fields. They come to our attention when people begin to refer to "the language of sports-casting," "the language of biophysics," etc. As we have seen, a grammatical sentence in a Sublanguage of English may not be grammatical in standard English even though the text in which the sentence occurs is still said to be "in English." When we speak of "the language as a whole" we include all such texts, thus it seems that a grammar of the language as a whole must describe all the sublanguages in it—certainly no mean task.

Many of the sentences of a Sublanguage of L are considered "standard L"; the percentage varies within each Sublanguage. And those sentences that are not so considered can be paraphrased in standard L (Check reservoir full ↔ Check to ensure that the reservoir is full). This suggests that the standard language may be useful in describing the way a Sublanguage fits into the language as a whole. Furthermore, sublanguages overlap and their interrelations form a part of the description of the language as a whole. A language is not simply a union of sublanguages, but a composite including many sublanguages related to varying extents lexically, syntactically and semantically. These relations are implied by statements like the following:

(i) In aeronautics the noun *dope* refers to a chemical compound used to coat fabrics employed in the construction of aircraft, whereas in pharmacology it may refer to narcotics.

(ii) The words *hammer*, *anvil*, and *stirrup* as used in the study of the ear are related metaphorically to these words as used in a smithy.

(iii) Instruction manuals in many fields employ a telegraphic style, often omitting the definite article in contexts where *it* is required in standard English.

(iv) Expression of emotion may be appropriate in a religious publication, but not in a journal of physics or math.

(v) The philosophy student's thesis was criticized for containing flip comments more appropriate to a term paper in freshman English.

(vi) English texts in various fields share the alphabet a, b, c, . . . , x, y, z but ∃ occurs in those dealing with mathematical logic, ə in phonology texts, etc.

Formalization of such relations may shed light on the role of sublanguages in the language as a whole.

Individual Sublanguage grammars are of independent interest for the purpose of information retrieval and automatic translation. A question which stands in need of more investigation is the extent to which corresponding sublanguages in different languages have similar characteristics. E.g., Kittredge claims ([5]) that variation in textual linking devices may be greater between two dissimilar sublanguages in the same language than between *two* corresponding sublanguages in different languages. If true, this is further evidence of the practicability of automatic translation between corresponding sublanguages in different languages. Of course one may point out individual characteristics of certain sublanguages which do not carry over to other languages. E.g., omission of the definite article in the texts described in section 2 does not occur in the French translation of these texts. Transfer rules are required to insert the appropriate form of the definite article in the French texts.

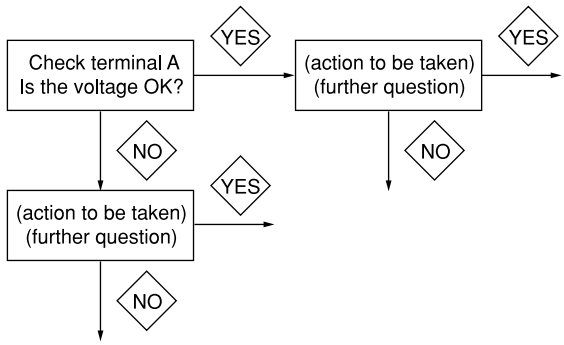
Within a given language there may be groups of sublanguages that have many characteristics in common. For example much technical writing in English differs more in vocabulary than in syntax. Thus it may be possible to construct a parser whose syntactic rules will suffice for a number of "technical" sublanguages, with only minor variations, even though there are considerable differences in vocabulary and in the semantic ranges of individual lexical items from one Sublanguage to another.

Other possibilities for further study of sublanguages not touched on here include phonological traits (e.g., in religious sermons), the growth of sublanguages along with scientific developments and cultural changes, and possible effects of usage within a Sublanguage on usage in other parts of the language.

Notes

1. These figures represent the stage of development at the end of 1978.

2. In certain related texts direct questions occur, but in well marked environments such as flow charts describing troubleshooting procedures:



References

[1] Chevalier, M., et al. (1978): TAUM-METEO: description du système, Groupe de recherche en traduction automatique, Université de Montréal.

[2] Downing, P. (1977): On the Creation and Use of English Compound Nouns, *Language*, vol. 53 (4).

[3] Harris, Z. (1968): *Mathematical Structures of Language*, John Wiley and Sons, New York.

[4] Isabelle, P., et al. (1978): TAUM-AVIATION: description d'un système de traduction automatisée de manuels 'entretien en aéronautique', COLING, Norway.

[5] Kittridge, R.: *Variation and Homogeneity of Sublanguages*, Chapter 4, this volume.

[6] Levi, J. N. (1978): *The Syntax and Semantics of Complex Nominals*, Academic Press, New York.

17
The Proper Place of Men and Machines in Language Translation

Martin Kay

Introduction

Important contributions to linguistics are likely to be more and more in the spirit of artificial intelligence. A manifesto for this point of view is out of place here. What is to the point is that translation embraces every facet of language while providing a task whose criteria of success, for all their problems, are remarkably well defined. In science in general, and artificial intelligence in particular, the proper role of computers is quite different from any they can play in engineering or any enterprise directed towards the fulfillment of immediate and practical needs. Here they are properly applied to what is not understood with the expectancy that, as much by their frequent and resounding failures as by anything else, they will illuminate the boundaries of our ignorance. Engineering pays heavily for the very failures that science can best profit from.

The need for translated texts will not be filled by a program of research that devotes all of its resources to a distant ideal, and linguists and computer experts will be denied the proper rewards of their labors if they must promise to reach the ideal by some specific time. A healthy climate for FAHQT will be one in which a variety of different though related goals are being pursued with equal vigor for the intellectual and practical benefits that they may bring.

There was a long period—for all I know, it is not yet over—in which the following comedy was acted out nightly in the bowels of an American government office with the aim of rendering foreign texts into English. Passages of innocent prose on which it was desired to effect this delicate and complex operation were subjected to a process of vivisection at the hands of an uncomprehending electronic monster that transformed them into stammering streams of verbal wreckage. These were then placed into only slightly more gentle hands for repair. But the damage had been done. Simple tools that would have done so much to make the repair work easier and more effective were not to be had presumably because of the voracious appetite of the monster, which left no resources for anything else. In fact, such remedies as could be brought to the tortured remains of these texts were administered with colored pencils on paper and the final copy was produced by the action of human fingers on the keys of a typewriter. In short, one step was singled out of a fairly long and complex process at which to perpetrate automation. The step chosen was by far the least well understood and quite obviously the least apt for this kind of treatment.

Government and bureaucracy may be imbued with a sad fatalism that forces it to look to the future as destined to repeat the follies of the past, but we can surely take a moment to wonder at the follies of the past and nostalgically to muse about what a kinder and more rational world would be like.

The case against machine translation as a solution to practical problems is overwhelming and has been made many times. I do not propose to repeat it in any detail here. It will, however, be worth a few words to make a *prima facie* case for the implausibility of practical machine translation if only so that the contrast with realistic approaches to the problem will be more striking. I will go on to outline what some of these might be. I shall make some specific proposals, but I should like it to be clearly understood that I do not believe that they represent the only course to follow. They are intended only to illustrate my main point, which is this: There is a great deal that computer scientists and linguists could contribute to the practical problem of producing translations, but, in their own interests as well as those of their customers, they should never be asked to provide an engineering solution to a problem that they only dimly understand. By doing only what can be done with absolute surety and reliability now and by going forward from there in short, carefully measured steps, very considerable gains can be virtually guaranteed to all concerned.

It is not difficult to convince oneself that fully automatic machine translation is no more a serious answer to any practical problem today than it ever

was. But to do so responsibly and scientifically requires examination of a lot of evidence and the careful design and performance of a number of experiments. It is clearly pure irresponsibility to attempt to assess any particular translation system on the basis of intuitive reactions to a so-called *demonstration* in which one examines what the line printer delivers and listens to ingenious attempts to explain the tenuous relationship this bears to, say, English. But it is reasonable and easy to consider what *prima facie* case can be made just on the basis of the advertising. This section contains two related arguments against the plausibility of machine translation as an industrial enterprise, one from the point of view of linguistics and the other from that of computer science.

Machine Translation and Linguistics

Let us look for a moment at a particular problem—one of the prodigious set that the designer of any machine-translation system has to face. Almost any member of the set would serve my purpose, so I will take one of the oldest and most hackneyed, namely how to choose a translation for a pronoun. To state it in this way is, of course, already a gross simplification because no translator or translation system worth its salt chooses translations for pronouns or, for that matter, any other isolated words. But the simplification will take nothing from the very elementary point I want to make.

Consider the following pair of sentences:

Since the dictionary is constructed on the basis of the text that is being processed, it need refer to only a small amount of context to resolve ambiguities.

Since the dictionary is constructed by a native speaker of the language, he need refer to only a small amount of context to resolve ambiguities.

Suppose that these are both translations from some other Indo-European language so that, in all probability, the underlined *it* and *he* correspond to the same word in the original; in French, the word would be *il*, for example. Now, it is entirely possible that automatic translation systems could be found that would get the translations of both sentences right. But it is almost inconceivable that one could be found that would get them right for the right reasons or that it would systematically solve problems of this kind correctly. If such a system did exist, we could not expect to find its designer in a gathering of this kind because he would surely be a person of such saintly modesty and so retiring a nature as to prevent his ever making his results known to others. He would die in poverty and obscurity. A person with normal human weaknesses who had the key to this problem could confidently be expected to claim the crown that linguistics is eager to bestow on him. Pronominal reference is, after all, among the most vexing problems in linguistics. Much the same can be said of innumerable other problems on whose solution the success of machine translation turns. If any of them had in fact been solved, we should not have to purchase an expensive system to find out about it and commercial or proprietary interest would not long hide it from us.

We are forced to one of two conclusions. Either some essentially *ad hoc* solution to these difficulties has been found and built into the systems that are offered for sale, or these systems do not really solve the problems at all. *Ad hoc* solutions tend to be based on case-by-case analyses of what linguists call *surface* phenomena, essentially strings of words, and on real or imagined statistical properties of particular styles of writing and domains of discourse. In scientific and technical texts, for example, one runs less risk of error by translating the French pronouns *il* and *elle* as *it*, rather than *he* and *she*, especially in contexts like *Il est possible que*.... In *Il est convaincu que* ..., on the other hand, *he* is a better bet. These facts are listed in the functional equivalent of a dictionary of *words in context* to which new entries are continually brought. The cash value of each new addition is slightly less than that of the one before as the contribution to the device as a whole slowly approaches its asymptote.

In fact, such little documentary evidence as the proprietors of past machine-translation systems have been prepared to release has typically pointed with evident pride to the great number of *ad hoc* devices that they contain and has made the incontestable point that any enhancement of the system in the future will require more and more and more of the same. I will come shortly to the question of what is wrong with engineering products that rest on *ad hoc* devices rather than sound theory.

Machine Translation and Computer Science

The *prima facie* case against operational machine translation from the linguistic point of view will be to the effect that there is unlikely to be adequate engineering where we know there is no adequate science. A parallel case can be made from the point of view of computer science, especially that part of it called *artificial intelligence*. To translate is to re-express in a second language what has been understood by read-

ing a text. Any purported solution to the problem that does not involve understanding in some sense is, at best, *ad hoc* and therefore subject to the linguistic objections already alluded to. A large part of the field of artificial intelligence is given over to building models on the basis of which to attempt some explication of this notion of understanding but no serious worker in this field has ever claimed to be able to provide the theoretical support required by any practical enterprise, least of all one so embracing as language translation.

There are also some points about past performance that deserve to be made from the computer scientist's point of view. There is, for example, the question of programming style and technique. The designers of machine translation systems have been intensely concerned with a property of their programs they call *efficiency*. Here is how the argument goes: These systems would not be required at all if there were not large quantities of text to be translated so that if one program took only slightly more computer time than another, it could soon involve a great deal of extra cost when put into operation. This is the main justification for the fact that there has been little or no use of higher-level programming languages. The lower-level assembly languages that have been used give the programmer direct access to the most basic facilities in his machine so that they cannot be less powerful than higher-level languages. Given a program in a higher-level language, it is almost always possible to produce a translation into assembly language that requires less machine time to run. This is not to say that it will not be a difficult, error-prone, and time-consuming operation to do so.

Against this obvious advantage of assembly languages must be set their equally obvious disadvantage, namely that they are arch-enemies of clarity and perspicuity. My claim will be that a program written in assembly language is much more likely to embody an *ad hoc* solution to a problem than one written in a higher-level language. This is only to be expected. Assembly languages give equal status to every detail in the specification of the program so that there is no way in which the overall plan that the program embodies can emerge. Consequently, a program that would seem simple in another language is almost guaranteed to look bewilderingly complex in assembly language. A program such as machine translation would require, one that would be complex by any imaginable standard, would be beyond imagination in assembly language. In programming, as in any kind of writing, the most complex ideas require the greatest clarity and skill for their exposition. But programming differs from everyday communication in that the languages available differ greatly in expressive power and the choice among them severely conditions the clarity that can be achieved. Every computer scientist is taught, but only comes truly to appreciate as a result of bitter experience, that programs are written for a human as well as a mechanical audience and the most important member of that audience is himself. A programmer who writes in assembly language is necessarily giving us less than his best at the highest possible price.

Efficiency, in computer programming, is itself a complex and subtle matter. It is true that it is affected by such issues as the language that a program is written in, but these effects are, at worst, linear. More realistically speaking, they are sublinear because a very large proportion of the time taken for any large program to run is accounted for by a very small proportion of the code. Standard practice, therefore, is to write a program in a language that displays its structure as clearly as possible and to rewrite carefully selected small portions of it in assembly language only when experience has clearly demonstrated that the effort involved in doing this would amply repay the effort.

Truly significant gains in efficiency invariably come from adjustments to the algorithm itself, that is, to the overall strategy that the program employs. Consider a simple example. The words of a text are to be looked up in a dictionary. There are a great many strategies that could be used, all of which would produce identical results for the same words and the same dictionary, but at very different cost in machine time. The dictionary could be searched from the top for each separate word in the text. Binary search, hashing, or one of the innumerable variants of these could be used. A method that has been popular with the designers of machine-translation systems is to sort the words of the text into alphabetical order so that a single pass through the resulting list and the dictionary locates all relevant entries. These are then sorted back into the order of the text.

Quite independently of the machine or the programming system used to implement them, these techniques can all be analyzed in terms of the way the time they take is related to the length of the text and the size of the dictionary. If there are m entries in the dictionary and n words in the text, if the dictionary is not ordered in any especially helpful way, and if almost all the words are in the dictionary, then the first method requires each of the n words in the text to be

compared with about half the words in the dictionary so that the time involved will vary with both m and n, in other words, it will be proportional to mn. Putting the dictionary in order by frequency could conceivably improve things to the point where the average word is found in the first $\log m$ entries, which would make the technique as efficient as binary search. A suitably chosen hashing scheme removes the effect of m altogether so that, at least from this point of view, this method is better than either of the others. The technique that requires sorting proceeds in three steps, two sorts and the comparison with the dictionary. The comparison with the dictionary is linear, but sorting n items takes on the order of $n \log n$ steps by the best known methods, which means that the time taken by the comparison is not significant. This is a simple classic case where the considerations determining the best solution are well known. In reality, the choice is often difficult and text-book solutions are not available.

There is a branch of computer science called *analysis of algorithms* that is devoted to the assertions of this kind that can be made about computational methods. What is important about such assertions is that they characterize the cost of a technique as *a function of* the data it will be applied to. Differences in the functions that characterize competing techniques are altogether more significant than the purely linear differences that programming languages and coding practice can affect. If techniques A, B, and C all achieve the same results when applied to an input of size δ, and the time taken by A varies with δ^2, B with δ, and C with $\log \delta$, then C is best and A is worst and, unless δ is very small, the implementation details are beside the point. If C takes 10 steps for a certain case, then B will take about 1000, and A, 1,000,000. When the differences are as great as these—and they often are—the cost of the individual steps is irrelevant.

Any program that purports to translate natural text must clearly be orders of magnitude more complex than one that simply looks words up in a dictionary. It will always be susceptible of improvement, at least in a theoretical sense, not only in the quality of the results it delivers but also in the efficiency of the algorithms it incorporates. To be continually improvable in this way, a program must be perspicuous and robust. It must be perspicuous so that there is never any doubt about the role that each of its parts plays in the overall structure and robust so that it can be changed in important ways without fear of damage. Perspicuity and robustness are clearly two sides of the same coin. They are the high ideals to which the art of programming is continually striving and which it never achieves.

The Statistical Defense

It is immediately clear why *ad hoc* solutions should be offensive to a scientist. His job is, in a sense, precisely to reveal as principled and orderly what had previously been *ad hoc*. But what we must attend to is whether these solutions should upset an engineer or someone whose primary concern is getting a job done and, if so, to what extent. Two arguments are commonly made for *ad hoc* solutions to the problems of machine translation. The first is a simple statistical claim that can be dismissed almost as easily as it can be stated. The second is what I shall refer to as the *sorcerer's apprentice* argument.

The statistical argument rests on the fact that something can be complex *and* subtle without the complexities and subtleties being spread uniformly through it. Linguistics requires of its practitioners remarkable virtuosity in constructing examples of problems such as no existing or proposed computer system could possibly solve. But the claim is that we do not have to solve them so long as they do not crop up very often. We may not have an algorithm that will identify the antecedent of a pronoun whenever a human reader could but, if it can devise a method that will identify it most of the time, that will be good enough.

An algorithm that works most of the time is, in fact, of very little use unless there is some automatic way of deciding when it is and when it is not working. If it were able to draw a proofreader's attention to all the cases of pronominal reference that were in doubt so that these, and only these, would have to be examined by a human reader, and if a high proportion of the cases were known to be correctly handled, then the utility of the technique would be clear. But the statistical argument is usually stated in the weaker form.

Suppose that a good, reliable translation of a text is required and that a computer program is available that translates pronouns correctly 90 percent of the time. If there were some way to tell which 10 percent of the pronouns had been wrongly translated, it would be sufficient to examine these to verify the correctness of the translation (ignoring, for simplicity, other possible sources of error). But since this cannot be done, 100 percent of the pronouns must be examined. To find a pronoun and check that it is correctly translated is expensive relative to making the correc-

tion. Therefore, it does not matter very much if the program is right 90, 99, 80, or 50 percent of the time. The amount of work that it leaves for the repairman is essentially the same. Somebody may claim, however implausibly, that 10 percent of pronouns occur in contexts where the translation is not crucial. This would be a useful thing to know just in case these were precisely the instances that the machine translated incorrectly but no such argument has been, or is likely to be, upheld.

The real situation is much worse because there is more to translation than pronouns. A great many decisions of essentially the same difficulty must be made in the course of translating a single sentence. If there is reason to expect each of them to be correct 90 percent of the time, there need only be seven of them in a stretch of text to reduce the expectation of translating it correctly to below 50 percent.

The moral is clear. The overall efficiency of a translation system, human or electronic, is directly related to its reliability. If it falls short of the acceptable standard, to *any degree whatsoever*, it might as well fail grossly because the burden it places on the proofreader will be very large, and not notably different in either case. The efficiency of a translation system, like any other, must be assessed over all its components, human and mechanical.

The Sorcerer's-Apprentice Defense

The sorcerer's-apprentice argument is to the effect that the kind of incomplete theory that linguists and computer scientists have been able to provide is often a worse base on which to build practical devices than no theory at all because the theory does not know when to stop. When a theory proposes questions about the data to which it can provide only partial answers, it is often better that the question should never have been asked.

Consider the following version of an often quoted sentence:

The man looked at the girl with the telescope.

It will be pointed out that this can be translated word-for-word into French, and innumerable other languages, and gives a perfectly adequate result. It is, of course, ambiguous in various ways because of the different roles that the prepositional phrase can play in the syntactic structure. But French admits exactly parallel ambiguities so that any effort spent trying to decide whether the girl had the telescope or the man had it and used it to see her with, is wasted. In fact, such an effort can serve only to jeopardize the translation because, if it results in any but a word-for-word translation of this sentence, there is an unnecessary risk that it will be wrong. On the other hand, if the sentence had been

The man looked at the girl with penetrating eyes.

the question of whose eyes were involved would suddenly have been important because no acceptable word-for-word translation is possible; we are forced to choose between *aux yeux* and *de ses yeux*. *Avec* is a good translation for *with* in neither case. What algorithm will tell a translator that this case needs analysis whereas the first one does not? Perhaps the absence of an article before *penetrating eyes* gives the clue. This would indicate that

He looked at the girl with affection.

requires analysis. Unfortunately for the argument, it does not.

The main problem with the sorcerer's-apprentice argument is that the decision that a sentence could be translated without analysis can only be made after the fact. Analysis shows that there is more than one interpretation of a sentence at some level and further analysis shows that there is a single translation that is compatible with each of them. In short, the algorithm required to decide when analysis is required would have to use the results of the very analysis it is designed to avoid.

What the sorcerer's-apprentice argument does suggest is that the process of translation should proceed in the following nondeterministic fashion. Whenever the information needed to make a choice reliably is not available, all possibilities should be followed up independently. Furthermore, when an essentially arbitrary choice must be made, say between a pair of synonymous words, these possibilities should also be held open. Under this policy, a given sentence of input would yield a family of sets of sentences in the target language. The members of each set are presumed equivalent and the sets are distinguished by the different patterns of decisions that led to their production. If, by happy chance, there is a sentence that belongs to every set in the family, then it is presumably the safest, and possibly even the best, translation. Consider, for example, the following somewhat contrived French sentence:

Ils signeront le document pourvu que leur gouvernement accepte.

Possible translations, classified by family, are

I

They will sign the document supplied that their government accepts.

They will sign the document furnished that their government accepts.

They will sign the document provided that their government accepts.

They are going to sign the document supplied that their government accepts.

They are going to sign the document furnished that their government accepts.

They are going to sign the document provided that their government accepts.

etc.

II

They will sign the document provided that their government accepts.

They will sign the document on condition that their government accepts.

They will sign the document only if their government accepts.

They are going to sign the document provided that their government accepts.

They are going to sign the document on condition that their government accepts.

They are going to sign the document only if their government accepts.

etc.

The two translations come from two quite different analyses of the original. It could be only as a result of a quite remarkable chance that a pair of interpretations as different as these should fall together. They would not have done so, for example, if *accepter* had been a verb that showed a difference between its indicative and subjunctive forms or if a feminine or a plural noun had taken the place of *document*. However, since the sentences involving the phrase *provided that* belong to both sets, the choice of a translation can be narrowed to them because this neutralizes the ambiguity.

This technique does not depend on there being a sentence that appears in every member of the family. Whenever a single sentence occurs in more than one set, they can be reduced to a single set containing only the intersection of the originals. There are optimal ways of choosing sets to conflate so as to reduce the choice that must eventually be made to a minimum. Furthermore, it is not difficult to devise extensions of the procedure. If the sets in the family of translations are labeled in some way for the places in the analysis where a decision was made in the course of their production, then the differences between pairs of translations can be ascribed to specific sets of decisions. If there is no translation that belongs to all the sets, then the number and the difficulty of the decisions that need to be made to make the choice can be minimized. This is a topic I shall return to. For the moment, the point to note is that the observation on which the sorcerer's-apprentice argument is based tends to maximize the amount of computation to be done—just the inverse of their original intent.

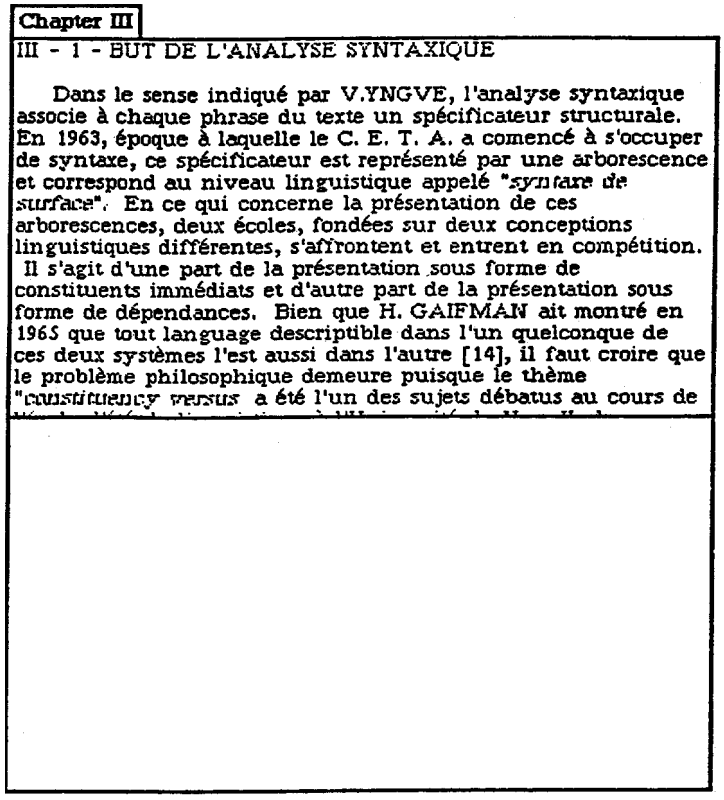
The Translator's Amanuensis

I come now to my proposal. I want to advocate an incremental approach to the problem of how machines should be used in language translation. The word *approach* can be taken in its original meaning as well as the one that has become so popular in modern technical jargon. I want to advocate a view of the problem in which machines are gradually, almost imperceptibly, allowed to take over certain functions in the overall translation process. First they will take over functions not essentially related to translation. Then, little by little, they will approach translation itself. The keynote will be modesty. At each stage, we will do only what we know we can do reliably. Little steps for little feet!

Text Editing

The easy way to prepare a piece of text is the way this one was prepared, that is, with a text-editing program on a computer. It does not matter whether it is done on a very small and personal computer that fits under the table in your office or on a large time-sharing machine, except that the latter is apt to be expensive. It matters very much that the design of the editor should be in the best possible taste, and it makes some difference whether the facilities include a screen that the writer can point at when he wishes to draw the program's attention to a particular place. People who have worked with bad editors soon retreat to the security of their typewriter or a pencil; anyone who has worked with a good one cannot be dragged away with a team of wild horses.

So, one thing to do would be to get a good editor and give it to your translators. If you could only do


Chapter III

III - 1 - BUT DE L'ANALYSE SYNTAXIQUE

Dans le sense indiqué par V.YNGVE, l'analyse syntaxique associe à chaque phrase du texte un spécificateur structurale. En 1963, époque à laquelle le C. E. T. A. a comencé à s'occuper de syntaxe, ce spécificateur est représenté par une arborescence et correspond au niveau linguistique appelé "*syntaxe de surface*". En ce qui concerne la présentation de ces arborescences, deux écoles, fondées sur deux conceptions linguistiques différentes, s'affrontent et entrent en compétition. Il s'agit d'une part de la présentation sous forme de constituents immédiats et d'autre part de la présentation sous forme de dépendances. Bien que H. GAIFMAN ait montré en 1965 que tout language descriptible dans l'un quelconque de ces deux systèmes l'est aussi dans l'autre [14], il faut croire que le problème philosophique demeure puisque le thème "*constituency versus*" a été l'un des sujets débatus au cours de

Figure 17.1
The initial display.

one thing, this would probably be the best. But you would do better to find an inventive computer scientist with good taste and to have him design a special editor which, in its earliest incarnations, would do little more than the program he would design for anyone else. But the design would be flexible and make provision for various kinds of extension. The kind of computer scientist I have in mind will expect to see the initial product in operation for a while before he makes detailed proposals for the extensions and he will probably want to see various alternative forms of each extension in use before any one is adopted.

Let us be specific. The device I am about to describe, which I call *The Translator's Amanuensis*, does not exist and probably never will. It is not the result of a careful program of design so that its details are ill specified, and what is specified I have invented only to illustrate the kind of avenue that seems most fruitful to follow and to avoid a long sequence of conditional sentences I should otherwise have to write.

Suppose that the translators are provided with a terminal consisting of a keyboard, a screen, and some way of pointing at individual words and letters. The display on the screen is divided into two windows. The text to be translated appears in the upper window and the translation will be composed in the bottom one. Figure 17.1 shows how the screen might appear before the translation process begins. Both windows behave in the same way. Using the pointing device, the translator can *select* a letter, word, sentence, line, or paragraph and, by pressing the appropriate key, cause some operation to be visited upon it.

There are various styles of work that a translator might adopt using this device. One that I shall pursue briefly here involves first copying the entire text to be translated into the bottom window. It thereby becomes, so to speak, the first draft of the translation. Little by little, words, phrases and sentences will be replaced by true translations until, in the end, little or nothing of the original remains in the bottom window. This somewhat unconventional procedure has the advantage of making it possible for the machine to maintain detailed linkages between the original and the translation so that it has a detailed idea of what corresponds to what.

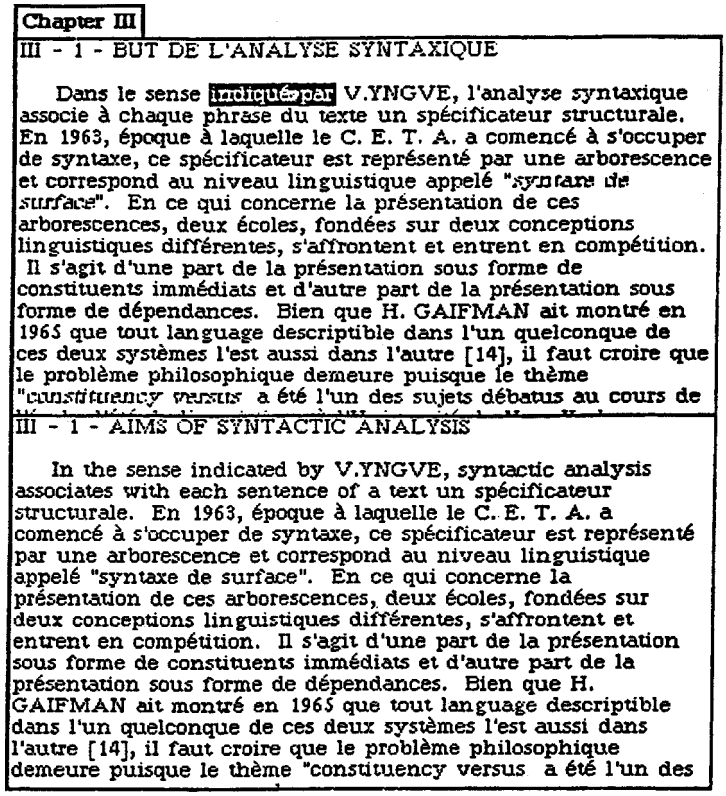


Figure 17.2
Selection.

In figure 17.2, the words *indicated by* have been selected because it has been decided that they constituted too literal a translation. The translator now gives the REPLACE command, say by striking R on the keyboard, and the selected word is replaced by a symbol showing that subsequent characters will be accepted as a new insertion at that place. In this case, the translator types *of* and the display is adjusted to show the amended text.

Translation Aids

This recital could be continued indefinitely. Basically, what I am describing is an editor of a kind that has become quite common. Now, let us consider how this device might be made to give special service to a translator. In line with the incremental approach, I will start simple. A relatively trivial addition would be a dictionary. The translator selects a word or sequence of words and gives a command to cause them to be looked up. In figure 17.3, the word *spécificateur* has been selected. When the lookup command is given, a new smaller window appears at a place indicted by the user. This new window gives the effect of overlaying some portion of the windows already present. In this case, the new window contains a deceptively simple dictionary entry for the selected word.

The simplicity of the dictionary entry is a feature of the system. We should think of the dictionary that the system has on file as being large and highly structured, growing on a daily basis as its deficiencies are revealed. To consult an entry, the user of the system is therefore provided with special tools. He is first shown only a gross summary of what the entry contains. By pointing to a subentry in that summary, he can obtain information on the next level of structure in a new window. The strange symbols following the words *Syntax* and *Semantics* in figure 17.3 represent text which will be included if the user points to them. The text that then appears may contain other instances of this symbol, and so on. The translator can thus cause the entry to develop in the direction indicated by the text on hand. At any time, he can return to a higher level by pointing at some part of the corresponding window that still remains exposed. If, in

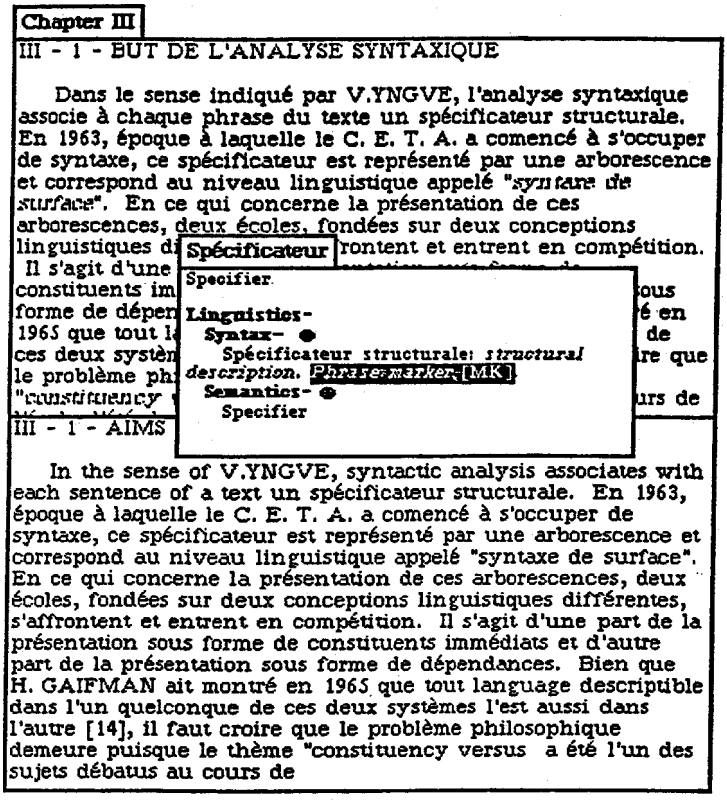


Figure 17.3
Looking up terms.

the course of translating a text, a word or phrase is looked up a second time, the display will show, not just the top-level entry, but the situation that was obtained when the same entry was consulted previously. This is on the theory that the same part of the dictionary is apt to be most relevant. The example given in the illustration is unrealistically simple; we must envisage many levels of structure and a greater investment of effort in the corresponding system design.

The translator can edit dictionary entries with the same commands that he uses for the translation itself. These amendments may be temporary, serving essentially as notes on the vocabulary of the particular document and the terminological decisions that have been made. They can also be more permanent, providing instant information to other translators with similar problems. Communication of this sort, across time as well as space, is one of the most crucial functions that computers can serve.

I take it that words selected for reference to the dictionary will not have to be in their citation form. The computer will be able to apply rules of morphological analysis to determine the proper dictionary heading for itself. While this is not trivial, it is one of the few parts of linguistic analysis that is well understood. Furthermore, it should be possible to look up compounded words and sequences suspected of constituting an idiom or fixed phrase.

The machine's dictionary can be used in a variety of ways. Suppose, for example, that a word is put in the local store—that part of the dictionary that persists only as long as this document is being worked on—if it occurs in the text significantly more frequently than statistics stored in the main dictionary indicate. A phrase will be noted if it occurs two or three times but is not recognized as an idiom or set phrase by the dictionary. By examining the contents of this store before embarking on the translation, a user may hope to get a preview of the difficulties ahead and to make some decisions in advance about how to treat them. These decisions, of course, will be recorded in the store itself. In the course of doing this or, indeed, for any reason whatever, the translator can call for a display of all the units in the text that contain a certain word, phrase, string of characters,

or whatever. After all, the most important reference to have when translating a text is the text itself.

If the piece of text to be translated next is anything but entirely straightforward, the translator might start by issuing a command causing the system to display anything in the store that might be relevant to it. This will bring to his attention decisions he made before the actual translation started, statistically significant words and phrases, and a record of anything that had attracted attention when it occurred before. Before going on, he can examine past and future fragments of text that contain similar material.

Most editing programs allow the writer to insert an arbitrary symbol of his own choosing at various places in the text and, at some later time, to cause all instances of that symbol to be replaced by some other word or symbol. This comes close to filling an important need that translators have. It turns out that a particularly vexing problem of technical translation is that of vocabulary control. That you should translate a technical term in one language by the proper technical term in the other language is important, but it is less important than that you should translate it always in the same way. One way to achieve this would be to make up a symbol, containing some otherwise unused character, and then to make replacements when the translation was complete. The device envisaged here goes further.

I suppose that the user of the system has available a special pair of brackets that he can insert in the text; in the examples they appear square and bold. They appear on his screen but will not be printed in the final translation. They are used as follows. If it is, for the moment, unclear how a word or technical phrase should be treated, the tentative translation is enclosed in these special brackets. They can be used in the translation itself or in dictionary entries. When the same word or phrase turns up again, the bracketed phrase is explicitly copied into the new position, thus maintaining an association among all the places where it is used. If the contents of such a pair of brackets is changed, the contents of all the others that are linked to it change automatically in the same way. Notice that this is a considerably finer instrument than the replacement facility of standard text-editing programs because the changes affect only those occurrences of a word or phrase that have been explicitly associated. Furthermore, if inflectional material belonging to one of these bracketed words or phrases is written in a standard, regular form outside the brackets—as in the case of the word *dependency* in figure 17.4—the appropriate form of the word can be constructed when the final version is settled on. Once again, this calls for the application of morphological rules, this time in the generative direction.

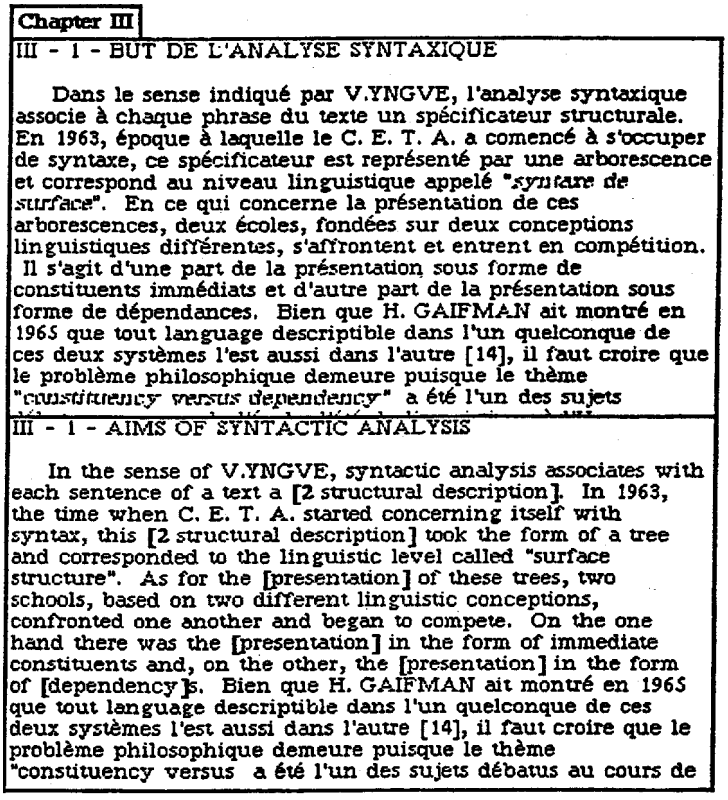
Figure 17.4 also illustrates another possible variant of this device. If more than one translation is being considered for a particular term, possibly because both are suggested in the dictionary, the fact is recorded by displaying a bold numeral just after the open bracket. If the translator points at this, the next possibility is taken, both in this and the other places in the text that are linked to it. In this way he can rapidly switch back and forth between variants without having to type.

Machine Translation

There is no early limit to the facilities that could, and probably should, be added to the translator's amanuensis. Rather than prolonging the rehearsal, let us look at where the process might end. I began by proposing an incremental approach to machine translation, so it is machine translation that must come at the bottom of the list. But, if it is to avoid the objections made by myself and others, it must be machine translation in a new form.

I propose that one of the options that should be offered to a user of the hypothetical system I have been describing, at a fairly early stage, be a command that will direct the program to translate the currently selected unit. What will happen when this command is given will be different at different stages of the system's development. But a user of the system will always be empowered to intervene in the translation process to the extent that he himself specifies. If he elects not to intervene at all, a piece of text purporting to translate the current unit will be displayed in the lower window of his screen. He will be able to edit this in any way he likes, just as post-editors have done in the past. Alternatively, he may ask to be consulted whenever the program is confronted with a decision of a specified type, when certain kinds of ambiguities are detected, or whatever. On these occasions, the system will put a question to the human translator. He may, for example, ask to be consulted on questions of pronominal reference.

The only difference between the translation facilities of the translator's amanuensis and previous machine-translation systems that can be seen from a user's point of view is that here the translator has his say while the translation is under way whereas previously he had to wait. If this scheme can be made


Chapter III

III - 1 - BUT DE L'ANALYSE SYNTAXIQUE

Dans le sense indiqué par V.YNGVE, l'analyse syntaxique associe à chaque phrase du texte un spécificateur structurale. En 1963, époque à laquelle le C. E. T. A. a comencé à s'occuper de syntaxe, ce spécificateur est représenté par une arborescence et correspond au niveau linguistique appelé "*syntaxe de surface*". En ce qui concerne la présentation de ces arborescences, deux écoles, fondées sur deux conceptions linguistiques différentes, s'affrontent et entrent en compétition. Il s'agit d'une part de la présentation sous forme de constituents immédiats et d'autre part de la présentation sous forme de dépendances. Bien que H. GAIFMAN ait montré en 1965 que tout language descriptible dans l'un quelconque de ces deux systèmes l'est aussi dans l'autre [14], il faut croire que le problème philosophique demeure puisque le thème "*constituency versus dependency*" a été l'un des sujets

III - 1 - AIMS OF SYNTACTIC ANALYSIS

In the sense of V.YNGVE, syntactic analysis associates with each sentence of a text a [2 structural description]. In 1963, the time when C. E. T. A. started concerning itself with syntax, this [2 structural description] took the form of a tree and corresponded to the linguistic level called "surface structure". As for the [presentation] of these trees, two schools, based on two different linguistic conceptions, confronted one another and began to compete. On the one hand there was the [presentation] in the form of immediate constituents and, on the other, the [presentation] in the form of [dependency]s. Bien que H. GAIFMAN ait montré en 1965 que tout language descriptible dans l'un quelconque de ces deux systèmes l'est aussi dans l'autre [14], il faut croire que le problème philosophique demeure puisque le thème "constituency versus a été l'un des sujets débatus au cours de

Figure 17.4
Morphology and lexical alternatives.

to work, as I claim it can, its many advantages are collectively overpowering.

The kind of translation device I am proposing will always be under the tight control of a human translator. It is there to help increase his productivity and not to supplant him. It will never resort to *ad hoc* measures that have not been explicitly sanctioned by him. In its normal and recommended mode of use, it will appeal to him rather than being forced back on unfounded guesses. After the system has been under development for a while, its users will either still be using it as a clerical aid, or they will be consigning considerable amounts of the actual translation work to it. The usage will remain mainly clerical only if the best efforts invested in the translation facilities failed to make them useful or economical. In other words, this system will certainly be able to undertake whatever present systems are able to undertake reliably and if that proves to be very little, the inference is clear. But there is reason to hope that it will undertake more. A system that never reached the stage of proposing translations would still be of inestimable value in automatically producing the final copy, looking up words and phrases faster and in a larger and constantly growing dictionary than would be possible any other way, in keeping notes about vocabulary usage and the like.

There are several important reasons to expect better performance of a system that allows human intervention as opposed to one that will brook no interference until all the damage has been done. First, the system is in a position to draw its human collaborator's attention to the matters most likely to need it. It is clearly important that he should give special attention to matters for which the designers of the system were unable to provide satisfactory algorithmic solutions. A wrong answer in these cases does nothing but mislead. It is far better that the labor and ingenuity spent on developing the machine's ability to make bad guesses should be employed more productively.

The second point is related. The decisions that have to be made in the course of translating a passage are rarely independent. The outcome of one decision typically determines whether certain other decisions will have to be faced at all. A wrong decision at the

beginning of such a chain leads the system to ask questions of the data before it that do not even make sense; whatever answer is given, it is bound to be wrong. Cascading errors of this kind are common in language processing. In the kind of system proposed here, they will not happen except when the human member of the team makes an error, or when he consigns too much to the machine. In the standard case, he will be consulted when the first decision in the chain is reached and will direct the machine along the right lines.

A third point concerns the machine's use of history. One of the most important facilities in the system is the one that keeps track of words and phrases that are used in some special way in the current text. It is a device that should probably be extended in a variety of ways to cover more than just vocabulary usage. By means of this, the translator is able to make a decision on the first occasion that a difficulty arises that will determine how both he and the machine treat it on all subsequent occasions. In other words, the man and the machine are collaborating to produce not only a translation of a text but also a device whose contribution to that translation is being constantly enhanced. A post-editor who changes something at the beginning of a translation must expect to make essentially the same change many more times before he finishes. This is not to say that a machine-translation program could not be devised that modified its behavior in the light of experience. However, such a system would be especially liable to the last objection made, namely that bad decisions made early lead to worse decisions later. It is bad enough that ill-founded decisions made early in the processing of a sentence should be allowed to engender other ones later; to extend this policy over an entire text is to invite disaster. The system proposed here will accumulate only experience of what was agreed upon between both human and mechanical members of the team, the mechanical always deferring to the human.

The translator's amanuensis will not run before it can walk. It will be called on only for that for which its masters have learned to trust it. It will not require constant infusions of new *ad hoc* devices that only expensive vendors can supply. It is a framework that will gracefully accommodate the future contributions that linguistics and computer science are able to make. One day it will be built because its very modesty assures its success. It is to be hoped that it will be built with taste by people who understand languages and computers well enough to know how little it is that they know.

18
Machine Translation as an Expert Task

Roderick L. Johnson and Peter Whitelock

1 Introduction

The case against fully automatic high quality machine translation (FAHQMT) has been well-canvassed in the literature ever since ALPAC. Although considerable progress in computational linguistics has been made since then, many of the major arguments against FAHQMT still hold (a good summary is given by Martin Kay (1980)).

It is not our intention to reopen the case for FAHQMT here. Rather, we contend that, accepting that FAHQMT is not possible in the current state of the art, it is both feasible and desirable to set up research and development programs in MT which can both produce results which will satisfy sponsors and provide an environment to support research directed towards bringing MT closer to the ultimate goal of FAHQMT.

This chapter describes the rationale and organization behind one such program, the UMIST English–Japanese MT project.

2 MT as Simulation of Translator Behavior

Since an ideal MT system will probably be expected by consumers of translations to exhibit the functional input-output behavior of an ideal human translator, it is not unreasonable to look to translators as a primary source of information about the problems of MT. Note that we are not saying here that an ideal MT system should necessarily be designed to model every aspect of the behavior of a human translator. We do believe, though, that important insights into the organization of MT systems can be gleaned from studying bow translators operate—and, more importantly, what kinds of knowledge translators use—when they do translation.

What this claim comes down to is the assertion that translation as currently practiced is a task entrusted to experts—the translators. What we try to do when we build an MT system is to incorporate all or part of the translator's expertise into a computer program. If we were able to characterize all of the expertise of the ideal translator in such a way that the characterization could be expressed as an executable computer program then, presumably, we would have attained FAHQMT.

Since we do not yet know how to achieve such a characterization, we look for a model which partitions translation knowledge in such a way as to maximize the efficiency of the human/machine collaboration, while at the same time facilitates transfer of responsibility from man to machine as our understanding of the act of translation improves.

3 Knowledge in Translation

We postulate that the professional (technical) translator has access to five distinct kinds of knowledge: target language (TL) knowledge; text type knowledge; source language (SL) knowledge; subject area ("real-world") knowledge; and contrastive knowledge.

We assume that the first four of these are not contentious: a translator must know both the language in which the translation is to be produced and the language in which the source text is written; (s)he should have sufficient command of the subject area and its associated stylistic conventions to make sense of the source text and to produce a target text which is acceptable to a subject expert TL speaker. It is worth noting here, in passing, that a good translator is normally expected to be able to compensate for lack of expertise in all of these except (typically) the first two, by appropriate use of external sources like native (SL) informants, monolingual subject specialists and reliable reference works. We shall return to this question in section 6.

The question of contrastive knowledge is a little more delicate. Many workers in MT advocate a two-stage translation model in which source and target texts are mediated by a linguistically neutral "interlingua." In such a model there is clearly no place for contrastive knowledge, or rather the relevant contrasts are between SL objects and interlingual objects,

on the one hand, and TL objects and interlingual objects on the other.

What we understand by contrastive knowledge is present typically in the so-called "transfer" models of translation, where both SL and TL components map between texts and "deep" representations or "interface structures" (IS). An SL (respectively, TL) IS, although it abstracts away from superficial idiosyncratic properties of texts, is still recognizably an SL (respectively, TL) representation. The role of contrastive knowledge—which in the limit case may be restricted to simple lexical equivalence—lies in determining how a given SL IS "translates" to the corresponding (set of) TL IS. We do not want to enter here into the debate on the relative merits of interlingual versus transfer organization in models of MT. As will transpire from the rest of the chapter, it makes little difference to our organizational proposals whether contrastive knowledge mediates between abstract SL and TL representations or between some SL linguistic and some interlingua. The main difference lies in the ease and consistency of formulation of the necessary knowledge by experts in the domain (linguists, lexicographers, and translators).

4 A Model of Translation

The basic model we propose, in over-simplified form, is the familiar transfer scheme shown in figure 18.1. The idea is that some analysis device A applies SL knowledge to a source text to produce a source internal structure IS; a transfer device T applies contrastive knowledge to the source IS to produce a target IS; and finally a synthesis device S applies TL knowledge to the target IS to produce a target text.

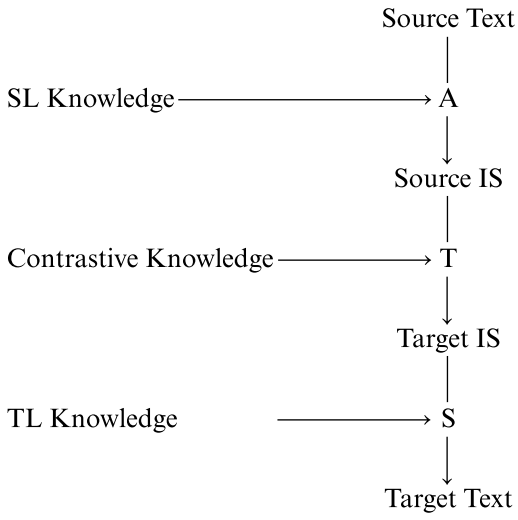


Figure 18.1
A model of translation.

In addition (not shown in the figure) all three of SL knowledge, contrastive knowledge, and TL knowledge may be enhanced by text-type knowledge.

In practice, as we all know, this model, even when enriched by text-type knowledge, is pathetically inadequate. For it even to have a chance of being useful, we should have to require that all of S, T, and A be total and functional. In practice, we know that this is unlikely ever to be the case with natural text.

Thus, we expect that the mapping computed by A (and its "inverse" S) will be many-to-many (one text may have many corresponding IS, many texts may have the same IS). Similarly, T is likely to be many-to-many, even if T only involves lexical substitution (consider *wall* vs. the Italian *muro|parete*, or *veal|calf* vs. the Italian *vitello*). Moreover both A and T will almost certainly in practice turn out to be partial (some tests will be ill-formed with respect to available SL knowledge, contrastive knowledge). The only thing we can reasonably enforce is that S should be total, by placing the requirement on T that it produce only well-formed IS representation.

There is, however, an important difference between the non-determinism inherent in A and T, on the one hand, and S on the other. If A, for a given text, produces multiple IS representations, then we assume that choice between them is not arbitrary and may have significant consequences for the correctness of the translation, although the available SL knowledge is inadequate to distinguish them. Similarly, the available contrastive knowledge may be inadequate to disambiguate multiple IS representations, although again the disambiguation may be important for the adequacy of the translation. On the other hand, once we have a target IS, the assumption is that all texts generable from that IS within the constraints of a given text type will be equivalent with respect to translation. If it is possible to derive more than one text from the IS in principle we do not need extra information to choose between the possibilities and the choice can be purely arbitrary.[1]

Thus we come up against two kinds of situations where linguistic knowledge in the system is potentially inadequate to meet the requirement of acceptable translation: when SL knowledge cannot disambiguate SL texts or contrastive knowledge available to the system fails to produce any result at all.

In some cases we can remedy such failures simply by adding to the available stock of linguistic knowl-

edge, as when the system fails to translate some text portion just because a word is missing from the dictionary. In many others there is no plausible linguistic solution; these are the cases where it is recognized that what is needed is an injection of subject-area or real-world knowledge.

Unfortunately, there does not exist, to our knowledge, any semantic or pragmatic theory which is sufficiently general and well-defined to allow incorporation into an MT model. Existing MT programs do what they can with what linguistics they have and leave the rest to human intervention. Our own view is also that practical MT should for the foreseeable future be a collaborative enterprise between human and machine. We want to claim, however, that current MT systems are generally not organized so as to make most efficient use of the human contribution. Moreover, we suggest that a well thought out design for an MT system should not only allow more efficient use of human resources but should also provide a useful research environment aimed at enhancing our understanding of the knowledge needed for translation.

In the next section, we look at conventional ways of organizing the man-machine partnership, before going on to our own design.

5 The Division of Labor in MT

Suppose we have a machine which can perform some part of the translation task, assisted by a human expert. There are essentially three points in the translation process when the human can intervene: after the machine has finished, while the machine is operating, or before the machine starts. It is worth remarking that once human intervention has ceased and the machine is left on its own, the machine's knowledge of what remains of the translation task must be complete.

We look briefly at each of the three possibilities in turn.

5.1 After—Post-editing

The safest way to organize man-machine cooperation in translation is to use a human post-editor to verify the output of an MT program, as is done in many large organizations using MT, especially where post-editing of work done by human translators is anyway the norm.

Post-editing is a highly skilled task; the post-editor needs to be an expert in:

- the subject area
- the target language
- the text-type
- contrastive knowledge.

In effect, the post-editor should be at least as skilled in all of these domains as the original translator. When the task of the translator is being done by a machine, it is not at all evident that we can claim that the machine is usefully extending expert capabilities to non-experts. At best, the computer is being used as a tool for the expert to increase productivity.

5.2 During—Interactive MT

A number of systems currently in use display the source text in the screen and provide facilities to allow the operator to build up a translation interactively, usually in a second window. Typically, the facilities provided include a window-oriented word-processor and on-line bilingual glossaries. In addition, such systems tend to offer an interactive "translation" mode, in which the machine attempts a sentence-by-sentence translation, pausing to prompt the operator to choose from among possible translation options; for example, the system might prompt:

Shall I translate "party" as
1. *partido*
2. *fiesta*

This way of working does not really differ from the post-editing scenario above. The possibility of interaction is only used to reduce the size of text fragments to be post-edited from full texts to sentence-sized units. Thus, although it appears to increase productivity (Hundt, 1982), it does not relieve the operator of any responsibility for any part of the translation task. The human end of the collaboration still needs to be carried out by an expert operator, who needs to possess all the expert skills of a translator.

5.3 Before—Pre-editing

In the pre-editing case there is at least some part of the translation task for which the machine is totally responsible (that part which happens after the last human intervention). Typically, in pre-editing environments, documents have to be specially drafted in a limited language using a restricted syntax and restricted vocabulary. The bargain is that the user guarantees only to submit input in the restricted language; the system guarantees that it will translate any valid text in that language.

The division of expertise here is quite different. Now the human needs only active, expert knowledge of the restricted language; all other aspects of translation expertise are supplied by the machine.

The neatness of this partition is somewhat illusory, however. The success of such an arrangement depends on being able to design a restricted language which ensures that all of the machine's inherent knowledge sources can operate infallibly: (passive) SL knowledge, subject-area knowledge, contrastive linguistic knowledge, text-type knowledge and (active) TL knowledge. As a consequence, these restricted languages tend to become so specialized and unnatural as to place unreasonable demands on the expertise of the pre-editor.

6 Distribution of Knowledge in Human and Machine Translation

None of these characteristics seems to us to offer a completely satisfactory framework for designing MT systems in such a way that they can be made to approximate more and more closely to the performance of an ideal translator.

To get closer to this goal, we look at the question of the use a human translator makes of available knowledge, with a view to finding a more productive basis for the sharing of expertise between man and machine.

A human translator is, first and foremost, a target language expert, as is evidenced by the practice of large organizations which require translators to translate only into their native language. It is rare for translators also to have expert knowledge of the subject area of the documents which they translate: they are normally expected to compensate for any deficiencies in their expertise by having extremely good contrastive knowledge and by consulting informed sources (reference works and/or subject types which they have to translate, since they largely bear the responsibility for the stylistic appropriateness of the translations they produce). Source language is also required, of course, but that knowledge need only be passive, and can be limited to experience of the written form in the relevant class of text types.

It is instructive to see how this use of knowledge compares to the presuppositions which seem to be built into the majority of commercial MT systems. In both the post-edited and the conventional interactive schemes, it appears that users expect to have to massage the machine's output to make it more acceptable stylistically. "Style clearly seems to be the main problem in post-editing" (Lavorel, 1982). This view is certainly not consistent with the idea of an MT system as a target language expert.

MT systems with only pre-editing come much closer to treating the machine as an expert translator. Where they differ from human translators is in placing strong, even perhaps unreasonable, requirements on the originators of documents as a means of circumventing their own deficiencies.

7 Towards More Productive Interaction Strategies

The model we propose is intermediate between the pre-editing and interactive styles of MT. If the machine is to behave functionally as far as possible like a human translator, then we would like to free the user from any need to know about the target language, so that the machine has to be a TL and a contrastive expert, as well as having text-type knowledge built in. On the other hand, while we anticipate that the system will be more or less deficient in knowledge of the user's SL and in subject-area knowledge, we assume that these deficiencies can be remedied in consultation with a (SL) monolingual operator. In terms of the model of section 4, we now have the picture in figure 18.2.

It is, of course, one thing to say that the system makes up for its own shortcomings by consulting the operator. It is quite another to determine when and how such consultation should take place. Being able to determine when to trigger an interaction depends on an awareness on the part of the system that there is something which it does not know. We can distinguish two such situations:

(a) the input is ill-formed with respect to either A (the analysis) or T (transfer);

(b) the input is ambiguous with respect to either A or T.

These two situations may occur, respectively, in cases where (a) A (respectively, T) is partial, or (b) A (respectively, T) is not functional.

Now we can (and should) arrange matters so that any construct produced by A can be transferred (i.e., T is total over the domain of outputs of A). This means that interactions triggered by ill-formed input (case (a)) can be localized within A only. We are not enthusiastic about attempts by the system to go it alone in "repairing" ill-formed input (see Arnold and Johnson (1984) for discussion), although this does not rule out use by the system of its own SL knowledge

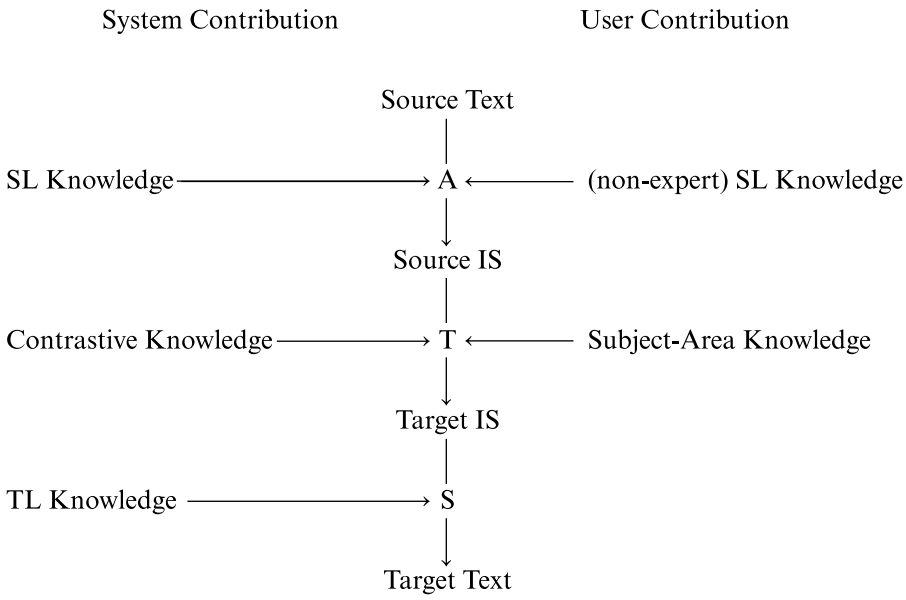


Figure 18.2
A modified model of translation.

to propose plausible reconstructions of the input as prompts to the user.

Case (b) is interesting, in that some apparent ambiguities in analysis may carry over to the TL (and so the system should not waste the user's time trying to resolve them). This observation suggests that this type of interaction should be handled as a part of transfer, utilizing contrastive information as criterion.

In cases of type (a), the system has a text fragment which it "knows" at analysis time it is unable to translate. The aim is thus to prompt the user to rephrase the input in a form which the analyzer can recognize. Thus the system must indirectly use its contrastive knowledge (knowledge of what it can translate) to extract from the user, who has extensive, but non-expert, SL knowledge, an acceptable formulation of the input text.

In type (b) cases, what has happened is that the purely linguistic knowledge available to the system is insufficient to distinguish between translationally distinct "readings" of the text. Hence the appeal to the user's "real-world" or subject-area knowledge to resolve the ambiguity.

We believe the approach advocated in this chapter to have two advantages over more orthodox MT systems design: it encourages a more efficient and productive sharing of expertise between man and machine; and it provides a useful framework for MT research by allowing the role of the machine to be extended incrementally on the basis of systematic experimentation within an operational environment. Most of the ideas are not original—indeed, the basis principles go back at least as far as Kay (1973). The same principles also seem to have been applied to MT by Tomita (1984). In our case the application domain is an experimental English–Japanese translator for technical documentation.

Acknowledgments

The work reported here is supported by International Computers Limited and the U.K. Science and Engineering Research Council under the Alvey program for research in Intelligent Knowledge Based Systems. We are grateful to ICL and to SERC, as well as to our colleagues in UMIST and the University of Sheffield, for their help and support in putting our ideas into practice.

Note

1. Actually, the situation with respect to T is not so clear-cut. Louis des Tombe has pointed out (personal communication) that, under certain very reasonable conditions, for any lexical item which is apparently unambiguous in the SL but ambiguous in the TL (e.g., the *wall* vs. *muro|parete* case above) if the ambiguity is resolvable with respect to the source IS then the information for resolving it should also be present in the target IS. Under these circumstances there is no reason why "disambiguation" should not be done by S,

provided the available TL knowledge is sufficiently precise to rule out the inappropriate case. On this view we should also have to accept that S may be partial.

References

Arnold, D. J., and R. L. Johnson. 1984. Robust processing in machine translation. In *Proceedings of COLING-84*, pp. 472–475.

Hundt, M. G. 1982. Working with the Weidner machine-aided translation systems. In Lawson (ed.) 1982, *Practical Experience in Machine Translation*. Amsterdam: North Holland. pp. 45–51.

Kay, M. 1973. The MIND system. In R. Rustin (ed.), *Natural Language Processing*. New York: Algorithmics Press, pp. 155–188.

Kay, M. 1980. *The Proper Place of Men and Machines in Language Translation*. Working paper CSL-80-11. Xerox PARC.

Lavorel, B. 1982. Experience in English-French post-editing. In V. Lawson (ed.) *Practical Experience of Machine Translation*. Amsterdam: North-Holland, pp. 105–109.

Tomita, M. 1984. Disambiguating grammatically ambiguous sentences by asking. In *Proceedings of COLING-84*, pp. 476–480.

19
Montague Grammar and Machine Translation

Jan Landsbergen

Introduction

In this paper I will examine the possibilities of using Montague Grammar for machine translation. I will discuss briefly the various ways in which this theory could be used, but most attention will be given to one actual application: the Rosetta translation system.[1] The paper is organized as follows. After a short introduction to Montague Grammar, its strong and weak points with respect to computer applications will be discussed. Then a syntactically powerful and computationally viable version of Montague Grammar, called M-grammar, will be described. Subsequently I will discuss various ways in which Montague Grammar may be used directly for machine translation and pay special attention to the problems that arise in these cases. Finally I will outline the isomorphic grammar approach to machine translation, followed in the Rosetta project, in which the compositionality principle of Montague Grammar plays an important role.

Montague Grammar

It is not possible to give in a few words a fair account of Montague Grammar and this holds in particular for its semantic power. In this section I will restrict myself to introducing some basic concepts and the corresponding terminology, which are needed for a good understanding of the rest of the paper. The terminology and the notation may deviate a little from "standard" Montague Grammar.

Montague's most important papers on language are "The Proper Treatment of Quantification" (1973), "Universal Grammar" (1970a), and "English as a Formal Language" (1970b). They have been collected together with other papers in Thomason (1974). A good introduction to the theory is Dowty et al. (1981). The 1973 "PTQ" paper, as it is usually called, is the best known and contains the most influential example of a Montague Grammar. The paper "Universal Grammar" describes the general algebraic framework (cf. Janssen 1986 for a better insight into and an elaboration of this framework). "English as a Formal Language" (EFL) is interesting because it shows how natural language can be interpreted directly, without intervention of a logical language.

The main characteristic of Montague Grammar is the attention that is given to semantics. Montague Grammars have to obey the compositionality principle, which says that the meaning of an expression is a function of the meaning of its parts. What the parts are has to be defined by the syntax, so the principle prescribes a close relation between syntax and semantics.

The syntax of a Montague Grammar specifies (1) a set of basic expressions and (2) a set of syntactic rules. The basic expressions are the smallest meaningful units, the syntactic rules define how larger phrases and ultimately sentences can be constructed, starting with the basic expressions. The rules are applied in a compositional ("bottom-up") way.

A simple example:

The basic expressions are: the noun *boy* and the verb *sleep*.

The rules are:

R_1: this rule is applicable to a noun, e.g. *boy*, and makes a definite plural noun phrase, by adding the article *the* and the suffix *-s*; e.g., *the boys*.

R_2: this rule is applicable to a noun phrase and a verb and makes a sentence with the NP as its subject, in the present progressive tense, e.g., *the boys are sleeping*.

The process of deriving a sentence from basic expressions by recursive application of rules can be made explicit in a syntactic derivation tree. In figure 19.1 an example of a syntactic derivation tree is given: it shows the derivation of the sentence *the boys are sleeping* according to the example grammar.

In Montague's example grammars the basic expressions and the expressions generated by the rules have a syntactic category, but no explicit internal structure, they are just symbol strings. Actually,

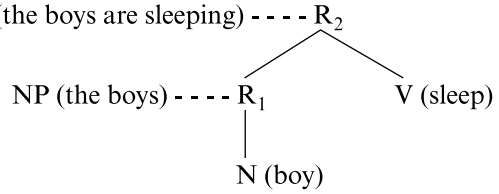


Figure 19.1

Montague used a version of categorial grammar. However, these restrictions are in general not considered essential properties of the theory. Already in the seventies Partee (1976) proposed an extension in which the rules operate on syntactic structures (or—equivalently—labeled bracketings) in which syntactic transformations may occur.

The semantic component of Montague Grammar assigns a semantic interpretation to the language as follows. First a semantic domain is defined, consisting of individual entities, truth values, special indices and functions defined in terms of these objects. Characteristic of Montague Grammar is the use of a special kind of indices, usually called "possible worlds". They are important for the power of the semantic system, which is often referred to as "possible-world semantics", but will not be discussed here.

The assignment of semantic values to expressions of the language can be done in two ways: directly and indirectly. In a direct interpretation (a method explored in the paper EFL) basic expressions and syntactic rules are immediately interpreted in terms of the semantic domain; each basic expression is associated with an object in the domain (e.g. an individual, a function from individuals to truth values, etc.) and with each rule an operation on objects in the domain (e.g. function application) is associated. The semantic value of an arbitrary expression is then defined with the help of the syntactic derivation tree. In parallel with the application of the syntactic rules the semantic operations associated with these rules are applied to the semantic values of their arguments, starting with the values of the basic expressions. The final result is the semantic value of the complete expression. So the process of derivation of the semantic value runs parallel with the syntactic derivation process and can be represented in a tree with the same geometry as the syntactic derivation tree, but which is labeled by names of semantic values and semantic operations. This representation, called semantic derivation tree, is introduced here because it will be useful in the sequel; it is not explicitly used by Montague. If we assume that the rules of our example grammar

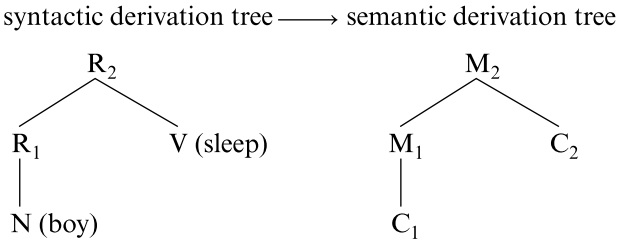


Figure 19.2

correspond to meaning rules, named M_1 and M_2, and the basic expressions to meanings with the names C_1 (for *boy*) and C_2 (for *sleep*), the relation between syntactic and semantic derivation tree is as in figure 19.2.

A simplified example:

C_1 is a property of individuals (equivalently: a set of individuals), i.e., the property "being a boy".

C_2 is also a property of individuals, i.e., the property "sleeping".

M_1 operates on a property P and yields the set of properties that all individuals with property P have, in this case the properties all boys have. (In this example it is assumed—wrongly—that "the+plural" can be interpreted as universal quantification.)

M_2 operates on a set of properties S and a property P and yields *true* if P is in S, else *false*.

So the semantic value of the sentence is *true* if the property of "sleeping" is a property that all boys have, else it is *false*.

The more usual way of assigning interpretations (pursued in PTQ) is the indirect one, which proceeds in two steps. First an expression of the language is translated into an expression of a logical language (in PTQ higher order intentional logic). Then the logical expression is assigned a semantic value by interpreting the logical language in the standard way.

The translation from natural language into logical language is defined in a similar—syntax-directed—way as the direct interpretation. For each basic expression its translation into the logic is given, each syntactic rule corresponds to a (possibly complex) operation on logical expressions. In parallel with the application of the syntactic rules the logical operations associated with these rules are applied to the logical expression associated with their arguments, starting with the logical expressions corresponding to the basic expressions.

The final result is the logical representation of the complete sentence. Note that in the indirect way of assigning interpretations, the form of the logical expressions themselves is not relevant; they are only a

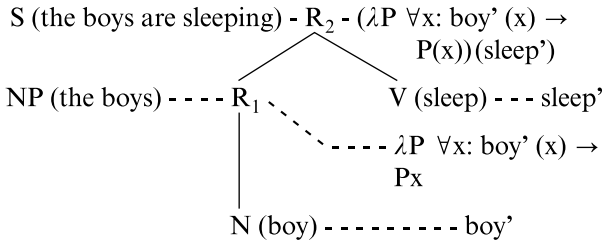


Figure 19.3

means to express in a convenient way the model theoretic interpretation.

In figure 19.3 I illustrate this process by showing in parallel the derivation of the sentence *the boys are sleeping* and of its (extensional) logical representation, but without further explanation. The derived logical expression for the complete sentence is equivalent to the reduced form: $\forall x$: boy$'(x) \rightarrow$ sleep$'(x)$.

Montague Grammar and Computer Applications

What are the strong and the weak points of Montague Grammar with regard to its use in computer applications that involve natural language processing?

Two important application areas in the field of natural language processing are natural language question-answering and machine translation. A strong point of Montague Grammar in these two areas is the attention that is given to semantics. In both application areas a sound semantic base is needed for determining what a correct answer or a correct translation is.

Another advantage of Montague Grammar in comparison with some other linguistic theories is its exactness and its "constructiveness". By "constructiveness" I mean that there is a clear step-by-step construction of phrases and—in parallel—of their meanings, thanks to the compositionality principle. Since for each rule both the syntactic and the semantic operation must be defined, the correctness of the rule can—to a large extent—be judged locally. This advantage is lost in a grammar with several syntactic levels, where the semantics is defined at the deepest level (whatever other virtues these levels may have). Local correctness criteria are important in the design of large systems in general and in particular in the design of large grammars.

A supposed weak point of Montague Grammar is that it treats only small fragments of language in a syntactically simplistic way. As for the fragmentariness, this is a consequence of exactness. Dealing with small—but nontrivial—fragments completely, in full detail is to be preferred—from the point of view of computer applications—to making interesting, but imprecise claims about natural languages in general. The syntactic simplicity of the framework is certainly a weak point, but it is more an incidental property of Montague's example grammars than an inherent property of the theory. The problem is not a lack of formal power, but a lack of linguistic power: the rules operate on strings and not on structured objects, e.g. syntactic trees. I have already referred to the syntactic extensions proposed by Partee (1976), and other work has been done in this direction, but nevertheless it is a correct observation that most workers in the field are primarily interested in semantics and less in syntax.

Another objection against Montague Grammar is that intentional logic and possible-world semantics are complicated and therefore hard to put to practical use in large systems. This is a correct observation. Montague needed the power of intentional logic to solve several difficult semantic problems, but these problems do not necessarily occur in all applications. For instance, in most data base question-answering systems a simple extensional semantics is sufficient. It is not in conflict with the spirit of Montague Grammar to use a simpler logic, as long as there is a compositional and model-theoretic semantics. The specific system of intentional logic may indeed be difficult, but model-theoretic semantics in itself is very easy to understand and to use; by imagining a particular interpretation it is possible to get a fast insight into the semantic correctness (and especially the incorrectness) of a particular rule or of a larger part of the grammar.

The most important obstacle to the application of Montague Grammar is that it is a purely generative framework. The theory defines how sentences and their meaning representations are generated in parallel, but it does not define how for a given sentence a meaning representation can be constructed effectively. This weakness can only be overcome by restricting in some way the class of possible Montague Grammars. This will be the topic of the next section. There I will define M-grammars, which are less powerful than unrestricted Montague Grammars from a purely formal point of view, but more powerful from a linguistic point of view, in the sense that the rules operate on structured objects instead of strings.

M-grammars

To my knowledge, two different ways of defining parsers for Montague Grammars have been de-

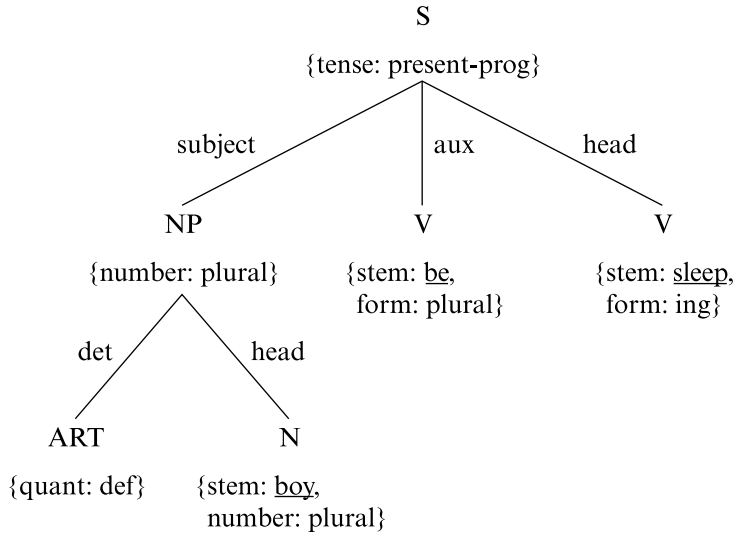


Figure 19.4

scribed: by Friedman and Warren (1978) and by Landsbergen (1981). The approaches differ strongly in what they consider to be a Montague Grammar. Friedman and Warren remained as close as possible to the PTQ grammar and designed a parser which can be characterized as a context-free parser with some specific extensions for phenomena falling outside the context-free framework, in particular the quantification rules. My own proposal defines a parser for a class of grammars, called M-grammars, which are syntactically more powerful and which are in accordance with Partee's transformational extensions (Partee 1976). Since 1981 a few changes in the definition of M-grammars have been made, of which the most important is the introduction of a separate morphological component. The new version is described in Landsbergen (1985). I will recapitulate it here briefly.

An M-grammar consists of three components: a syntactic component, a morphological component and a semantic component.

The syntactic component of an M-grammar defines a set of surface trees of sentences. The specific kind of surface trees generated by M-grammars—and the intermediate results—are called S-trees. An S-tree is an ordered tree of which the nodes are labeled by syntactic categories and attribute-value pairs and of which the edges are labeled by syntactic relations.

Formally, an S-tree t is an object of the form

$N\,[r_1/t_1, \ldots, r_n/t_n]$, ($n$ Š O)

with $N = C\ \{a_1\colon v_1, \ldots, a_k\colon v_k\}$

where

N is a node,

$t_1, \ldots, t_n$ are S-trees, the immediate constituents of t,

$r_1, \ldots, r_n$ are syntactic relations, between t and its constituents (if $n = 0$, t is a terminal S-tree)

C is a syntactic category,

$a_1, \ldots, a_k$ are attributes,

$v_1, \ldots, v_k$ are values of these attributes.

An example of an S-tree in the more familiar graphical representation is given in figure 19.4. It is a simplified—and unrealistic—example of a surface tree, for the sentence *the boys are sleeping*.

In the sequel I will often use an abbreviated notation, as in the following problem:

S (the boys are sleeping)

The leaves of an S-tree correspond to words. For example, the terminal node

N {stem: *boy*, number: plural}

corresponds to *boys*. This relation between terminal nodes and words as symbol strings is defined by the morphological component.

An M-grammar defines a language (in this case a set of surface trees) in the same way as a Montague Grammar, i.e., by specifying a set of basic expressions and a set of syntactic rules. But here the basic expressions are S-trees (in general S-trees consisting of one node) and the rules are defined for S-trees as arguments and yield S-trees as their results.

The derivation process of a surface tree from basic S-trees by application of rules can be represented by a syntactic derivation tree in the way described earlier. If we reformulate the example grammar of the previous section in terms of S-trees, the syntactic derivation tree of *the boys are sleeping* (i.e., of its surface tree) is the same as in figure 19.1.

In principle all rules of an M-grammar have "transformational power": they can perform fairly arbitrary operations on S-trees. However, this power is restricted by three conditions that M-grammars have to obey in order to make effective parsing possible: the reversibility condition, the measure condition, and the surface syntax condition. I will describe them here informally (cf. Landsbergen 1985 for more precise definitions).

The reversibility condition states that a rule should not only define a compositional ("generative") function (with a tuple of S-trees as argument and an S-tree as result), but also an analytical function (which operates on an S-tree and yields a tuple of S-trees). The compositional and the analytical function should be each other's reverse (the term *reverse* is used instead of *inverse*, because a rule produces a *set* of results, possibly the empty set, if the rule is not applicable). If the compositional function is applied to a tuple $(t_1, \ldots, t_n)$ and t is in the set of results, then application of the analytical function to t must yield a finite set containing the tuple $(t_1, \ldots, t_n)$, and vice versa.

Given a set of basic S-trees and a set of reversible rules, two functions, M-PARSER and M-GENERATOR, can be defined:

M-GENERATOR operates on an arbitrary syntactic derivation tree (i.e., an arbitrary tree labeled by rules and basic expressions) and yields a set of S-trees, by applying the compositional versions of the rules in the derivation tree, in a "bottom-up" way. The resulting set may be empty if some rule is not applicable.

M-PARSER operates on an arbitrary S-tree. It tries to apply the analytical versions of the rules in a "top-down" way until it arrives at basic S-trees. If this is successful, the result is a syntactic derivation tree (more than one derivation tree in case of ambiguities; the empty set if the input was not a correct S-tree).

M-GENERATOR and M-PARSER are each other's reverse: they define the same relation between S-trees and derivation trees.

In order to guarantee that M-PARSER is a computable function, an M-grammar has to obey the measure condition. It says: there is a measure on S-trees (a function from S-trees to integers, with a minimum) such that application of an analytical rule to an S-tree t yields S-trees smaller than t with respect to this measure. An example of a measure is the number of nodes in an S-tree, but in practice more subtle measures are needed. Thanks to the measure condition, application of M-PARSER always ends after a finite number of rule applications.

As it is our purpose to generate and analyze sentences, not surface trees, additional functions are needed. In the generative direction this is no problem: a function LEAVES can be defined which yields the sequence of leaves (the terminal S-trees) of an S-tree. For analysis purposes we need the third condition on M-grammars, the surface syntax condition. It says that for each M-grammar a set of "surface rules" must exist which define for each sentence a finite set of surface trees of which the set of correct surface trees is a subset. So this surface syntax has to be "weaker" than the real syntax and the surface rules can be simpler than the actual syntactic rules. A surface rule is applied in a bottom-up way to a sequence of S-trees; if it is applicable, the result is an S-tree with a new top node and with the input sequence of S-trees as its immediate constituents. Thanks to this, conventional parsing strategies can be used for the application of the surface rules, e.g., a variant of the CKY or the Earley Parser. The function applied by the parser is called S-PARSER.

The morphological component of an M-grammar relates terminal S-trees to actual words, symbol strings. It makes use of a dictionary and of various kinds of morphological rules, not to be discussed here. The morphological component defines two functions:

A-MORPH converts words into (sets of) terminal S-trees.

G-MORPH converts terminal S-trees into (sets of) words.

A-MORPH and G-MORPH are each other's reverse.

The syntactic component and the morphological component together define a function SYNTACTIC ANALYSIS and a function SYNTACTIC GENERATION, which are each other's reverse. The function SYNTACTIC ANALYSIS is the composition of A-MORPH, S-PARSER and M-PARSER, the function SYNTACTIC GENERATION is the composition of M-GENERATOR, LEAVES and G-MORPH. In figure 19.5 the two functions are shown with example

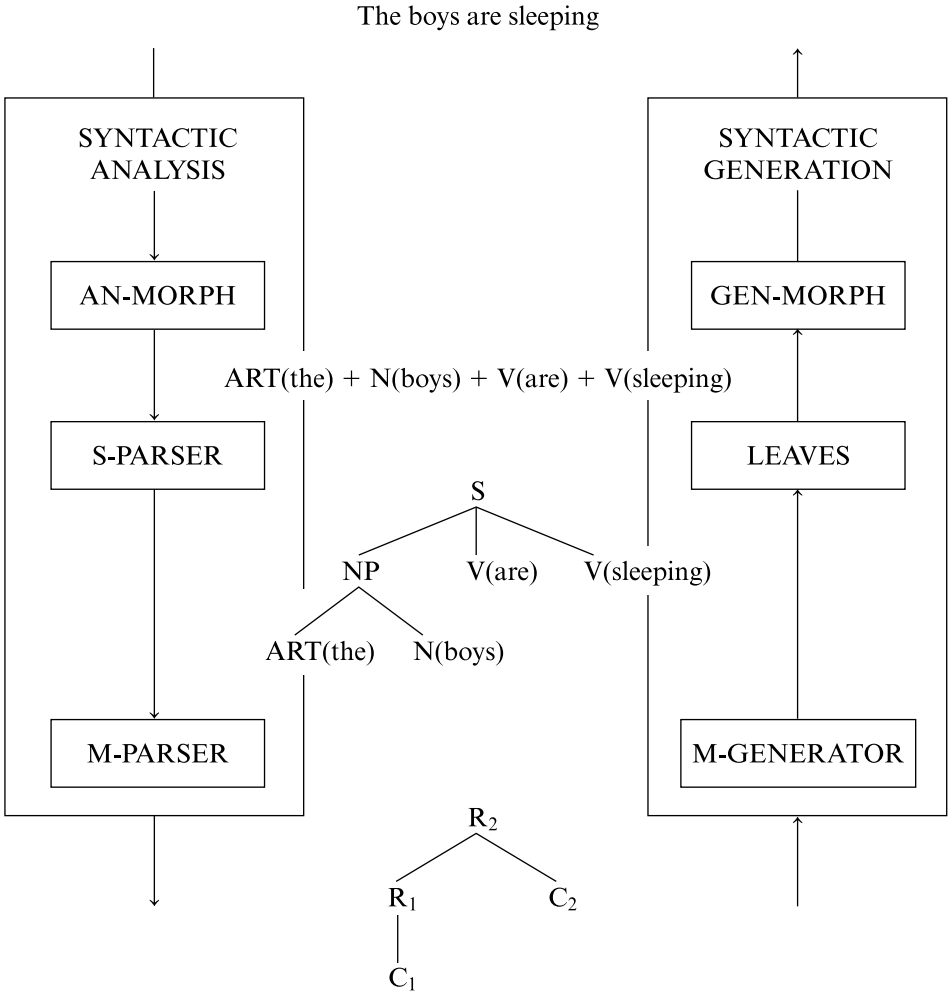


Figure 19.5

expressions. Note that the examples are a bit misleading as they suggest that these functions always give a unique result, which is the case for our example grammar, but not in general.

The semantic component of an M-grammar defines for each syntactic rule a "meaning rule" and for each basic expression a set of "basic meanings". As it depends on the application what the most appropriate way is to express these meanings—in an intentional logic, in an extensional logic or in some other way—this is left open here. A minor difference from standard Montague Grammar is that in an M-grammar a basic expression may have more than one meaning. This has the practical advantage that during analysis purely semantic word ambiguities can be "postponed" until after the syntactic analysis.

Montague Grammar and Machine Translation

I arrive now at the central topic of my paper: the use of Montague Grammar in translation systems. In the previous section I have defined M-grammars, syntactically powerful versions of Montague Grammars, for which an effective analysis procedure can be defined. In what way can they be used in a translation system? In order to be able to discuss the application of a linguistic theory in a translation system, I assume that in such a system the linguistic aspects can be clearly separated from the other aspects (e.g. the use of extralinguistic information, robustness measures, etc.). Then it is possible in principle to consider a "stripped" system that makes use of linguistic information only. In addition I restrict the discussion to systems that translate isolated sentences. Such

systems are in general not able to translate sentences unambiguously, but they define a set of possible translations. I define the function F-PTR as the function that operates on a sentence of the source language and yields the set of possible translations into the target language. F-PTR has the property that it is reversible: if s' is a possible translation of s, then s is a possible translation of s'.

s' in F-PTR(s) $\leftrightarrow$ s in F-PTR$'$(s')

The "correct" or "best" translation of s (chosen on the basis of extra-linguistic information) should be an element of the set F-PTR(s). Obviously, the function that yields this best translation is not reversible.

I would like to impose the following requirements on such a "possible translation" system.

1. It must be defined clearly what are correct sentences of the source language (SL) and the target language (TL). In other words, the system must be based on explicit grammars of SL and TL.

2. The translation function F-PTR must be defined in such a way that correct sentences of SL are translated into correct sentences of TL.

For me these requirements define the domain in which a theoretical discussion on machine translation makes sense. It is hard to compare—on a theoretical level—translation systems that do not obey them or at least try to obey them.

3. There must be some definition of the information that has to be conveyed during translation. Only if there is a clear definition of information content that a sentence and its translation should have in common, is it possible to evaluate a translation system in this respect. Unfortunately, there appears to be no theory of translation that offers a satisfactory definition.

The obvious way to use Montague Grammar (i.e., M-grammar or some other analyzable version) in a "possible translation" system appears to be the following. Define a Montague Grammar for the source language and for the target language. From these grammars analysis and generation components are derived. Then we extend the analysis with a component which translates a syntactic derivation tree into the logic according to the semantic component of the grammar. The generation component is extended with a component which performs the reverse function. So in this approach Intentional Logic is used as an interlingua. This type of system is outlined in figure 19.6.

This approach obeys the three requirements: a correct sentence of SL is translated into a correct sentence of TL according to explicit grammars and the information that is conveyed is the meaning in the model-theoretical sense. At first sight this is a very attractive method. It has the additional advantage that knowledge of the world can in principle be for-

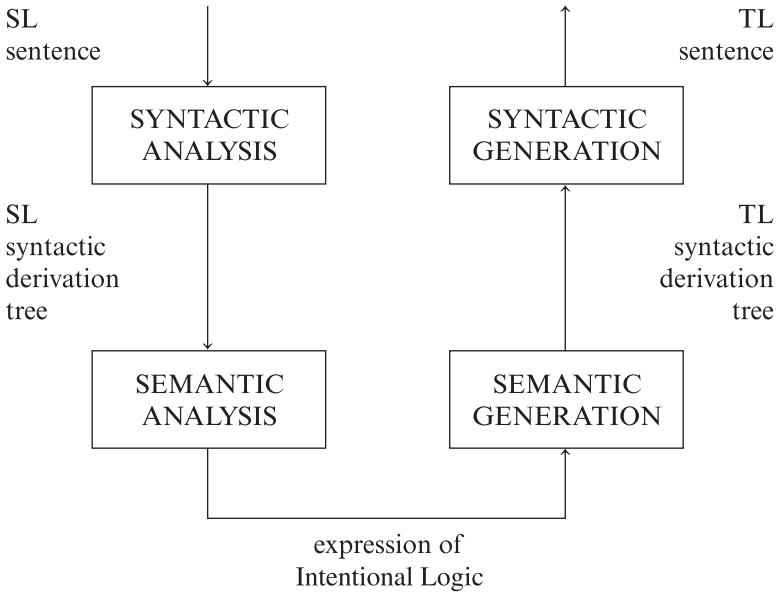


Figure 19.6

mulated in the same logical language as the interlingua, that inferences can be made, etc. I think that long-term research along these lines would be very useful. But in the Rosetta project we have chosen a different approach. Why? Because of the following problems with intentional logic as an interlingua.

1. Montague Grammar has been successful in defining the semantics of a number of natural language constructions, but a lot of work has to be done yet. For translation purposes it is in general not necessary to define in detail what a certain term or construct means, it is sufficient to know that a term or construct of one language means the same as a term or construct of another language. For example, the semantics of belief-sentences may be a problem, but the translation of the verb *believe* into the Dutch *geloven* is probably not at all problematic. This is not really a fundamental objection against the use of some kind of intentional logic. The problem is mainly that there is a discrepancy between the actual research in the field of Montague Grammar (directed to a detailed semantic analysis, for small fragments) and what is needed for machine translation (a fairly superficial analysis, with a wide coverage).

2. The second problem is more fundamental. In this approach the information that is conveyed during translation is the meaning in the model-theoretic sense. This is a nice basis for machine translation and certainly preferable to a purely syntactic approach, but there is other information to be conveyed as well, e.g., information on pragmatic and stylistic aspects. In general it seems to be wise to stay as close to the original form as possible (in some sense of the word "form"). Intentional logic is not adequate for carrying this information. One might object that the form of the logical expression expresses information about the form of the sentence too, and this is correct to a certain extent, but making use of the form of logical expressions is in fact in conflict with the spirit of Montague Grammar. As I already mentioned in the introduction, the logical expressions are only a way to define the model-theoretic meaning, their form is not relevant.

3. The third problem is the most delicate one: Montague Grammars translate natural languages into a subset of intentional logic. There is no guarantee that two Montague Grammars for two languages map them onto the same subset. In figure 19.7 the situation is sketched. The grammar of SL maps onto a subset IL_1 of IL. The grammar of TL maps onto a subset

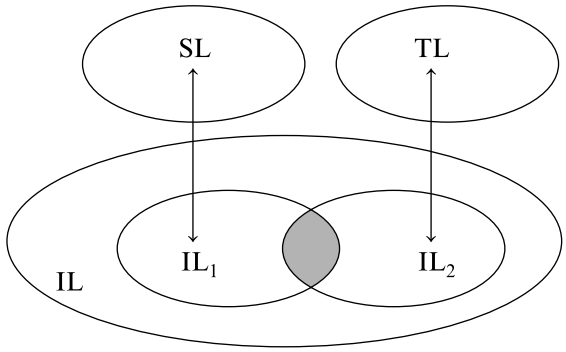


Figure 19.7

IL_2, and consequently the generation component based on this grammar is only applicable to expressions of IL_2. So translation is only possible for the sentences that are mapped onto the intersection of IL_1 and IL_2.

Notice that there is no independent definition of IL_1 and IL_2. They are only defined indirectly by the mappings that follow from the grammars of SL and TL. Therefore it is very difficult to get to grips with this problem. For solving it, it is not sufficient that the terms of IL_1 and IL_2 are the same, but in addition sentences that are to be translated into each other should get exactly the same logical structure and not just equivalent logical structures.

This "subset problem" arises in some guise in all systems—both interlingual and transfer systems—that translate via deep structures of some kind. In general it is not possible to define the translation for all "possible" deep structures (many of them will not correspond to any sentence at all), but on the other hand it is not possible to characterize what the subset of relevant deep structures is and to guarantee their translation. (Of course this problem does not arise in systems where the correct translation operations cannot be distinguished from the robustness measures.) The only fundamental way to solve this problem appears to be that the grammars of SL and TL are not developed independently, but in close cooperation. This possibility will be exploited in the next section, but will be left out of consideration here.

There are various other ways in which Montague Grammars can be used for machine translation. One of them is to make a transfer system at the level of the intentional logic. In terms of figure 19.7 the transfer component has to translate from IL_1 into IL_2. Godden (1981) has done work along these lines for Thai to English, making use of Friedman and Warren's

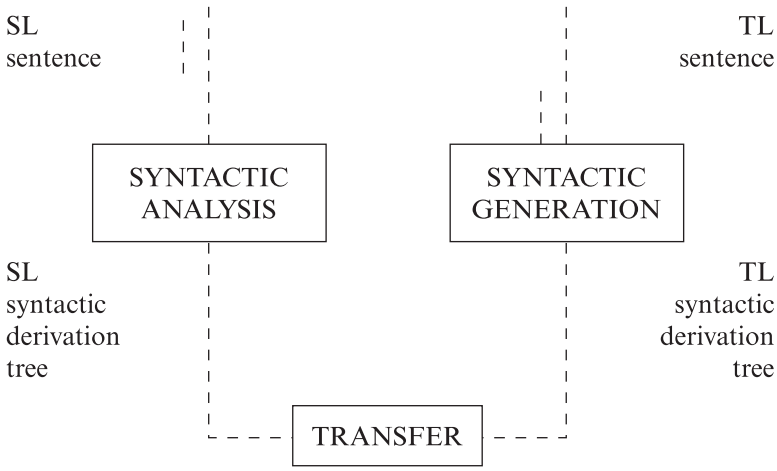


Figure 19.8

parser. The transfer rules have the status of meaning postulates, which gives them a sound semantic foundation. This is very interesting, but has only been worked out for the small fragment grammar of PTQ and does not appear to be easily extensible to larger fragments. Godden wrote in fact a PTQ-like grammar for Thai (i.e., the grammars for the two languages have not been written independently of each other) and added transfer rules for the small set of discrepancies between this grammar and the English PTQ grammar. Apart from the problem of the growing set of discrepancies for larger grammars (which ultimately comes down to the earlier-mentioned figure 9.3), figures 9.1 and 9.2 with regard to the use of intentional logic in machine translation are valid here too.

Another possibility of basing a translation system on Montague Grammar is to design a transfer system as outlined in figure 19.8 with transfer at the level of syntactic derivation trees.

In this approach there is an analysis component based on a grammar of SL and a generation component based on a grammar of TL; the transfer component converts syntactic derivation trees of SL into syntactic derivation trees of TL. In the most general version of this approach the transfer rules would convert arbitrary parts of SL derivation trees into arbitrary parts of TL derivation trees. Figures 9.1 and 9.2 do not arise here, as intentional logic is not used explicitly. However figure 9.3, the subset problem, returns here in a different form. The point is that the rules of the TL derivation tree that is yielded by the transfer component need not be applicable.

A different type of Montague-based transfer system is described by Nishida and Doshita (1982). In this system the transfer component converts the logical expression yielded by the analysis component (of which the terms are source language dependent) into a function-argument structure of which the application (in the generation component) yields target language expressions. There is no separate grammar of the target language in this approach.

I discussed the various Montague-based approaches under the assumption that the grammars of source language and target language are developed independently. Some of the problems are alleviated or disappear completely if these grammars are coordinated in some way. One, rather drastic, way of doing this will be discussed in the next section.

Isomorphic M-grammars

After the introduction of M-grammars, compositional grammars that can be used for both analysis and generation, only a relatively small, but essential, step has to be made to arrive at the isomorphic grammar approach. This step is that the grammars of the various languages are not developed independently, but more or less in parallel and are attuned to each other as follows.

For each basic expression in one language there must be at least one corresponding basic expression in the other language with the same meaning. For each syntactic rule in one language there must be at least one corresponding syntactic rule in the other language with the same meaning operation. Grammars that are attuned in this way are called isomorphic grammars, if the rules obey applicability conditions to which I will come back later.

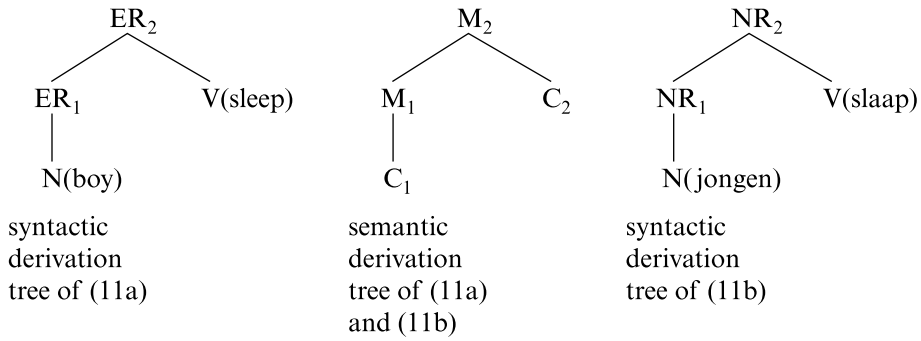


Figure 19.9

Given two isomorphic grammars, the translation relation is—informally—defined as follows: two sentences are translations of each other if they are derived from corresponding basic expressions by application of corresponding rules.

Before giving more precise definitions, I will give a simple example of isomorphic grammars for English and Dutch, in the table below. The grammar is the same as the one described before. In the middle column of the table the names of the basic meanings and meaning rules that the two grammars share are given. The grammars define a translation relation between sentences (a) and (b).

DUTCH		ENGLISH
basic expressions	basic meanings	basic expressions
N (jongen)	C_1	N (boy)
V (slaap)	C_2	V (sleep)
syntactic rules	meaning rules	syntactic rules
NR_1:	M1	ER_1:
N (jongen) → NP (de jongens)		N (boy) → NP (the boys)
NR_2	M_2	ER_2
NP (de jongens) + V (slaap) → S (de jongens slapen)		NP (the boys) + V (sleep) → S (the boys are sleeping)

(a) The boys are sleeping.
(b) De jongens slapen.

In the example grammar I use the abbreviated notations for S-trees; the rules are characterized by means of an example application.

Note that the relation between basic expressions of Dutch and English need not be one-to-one, although the example may suggest this. For each basic meaning there is a set of basic expressions in each language. The same holds for the rules. For example, NR_2 might also correspond to a rule ER_3, which generates a sentence in the simple present tense. Then the grammars would also define a possible translation relation between *de jongens slapen* and *the boys sleep*.

The definition of the translation relation given above can be reformulated more precisely as follows. Two sentences are each other's translation, if they have the same semantic derivation tree, i.e., if they have syntactic derivation trees with the same geometry, of which the nodes are labeled by corresponding rules and basic expressions. The syntactic derivation trees of the example sentences and their semantic derivation tree are given in figure 19.9.

There are several possible ways of using isomorphic grammars in a translation system; one of them is a transfer system like the one sketched in figure 19.8. The global design is the same, but the difference is that the TRANSFER component is now much simpler. The syntactic derivation tree of the source language can be converted into a derivation tree of the target language by a straightforward node-by-node transfer of basic expressions and rules.

Here I will discuss another possibility: the use of semantic derivation trees as interlingual expressions. This lies at hand, since a semantic derivation tree is exactly what translations have in common according to our definitions. In the section on M-grammars I described how a function SYNTACTIC ANALYSIS and a function SYNTACTIC GENERATION can be defined on the basis of the syntactic and the morphological component of an M-grammar. The semantic component of an M-grammar relates basic expressions to basic meanings and syntactic rules to meaning rules. On this basis two additional functions can be defined:

A-TRANSFER applies to a syntactic derivation tree and yields the set of corresponding semantic derivation trees.

G-TRANSFER applies to a semantic derivation tree and yields the set of corresponding syntactic derivation trees.

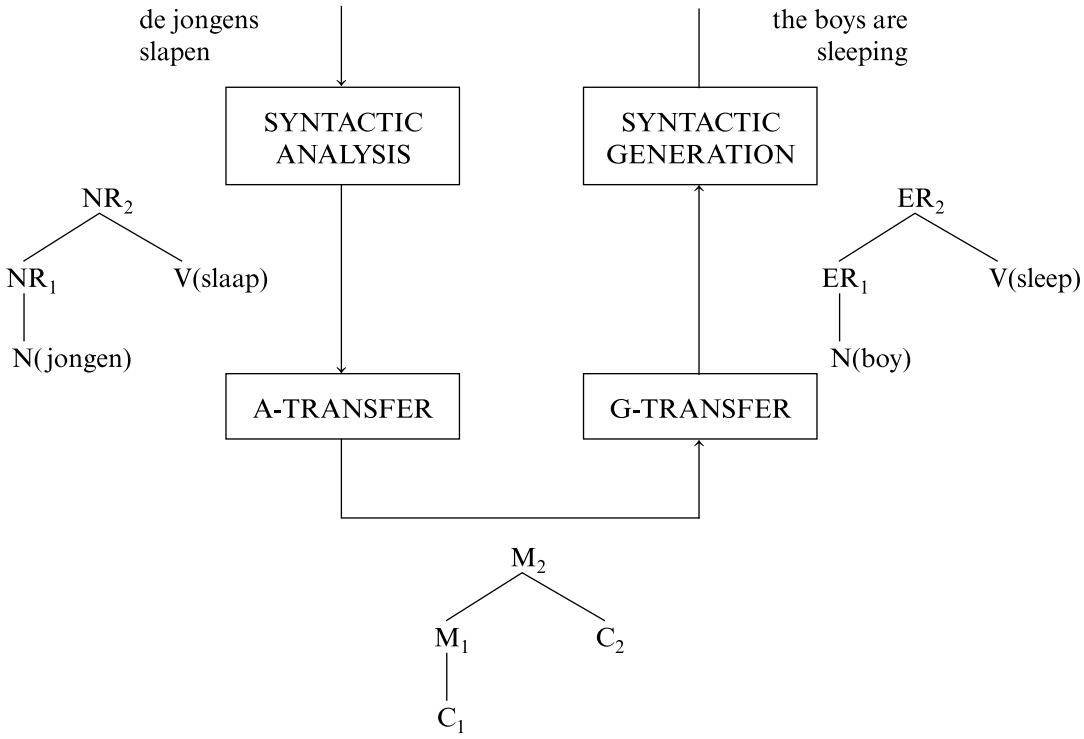


Figure 19.10

Both A-TRANSFER and G-TRANSFER are simple functions, defined in terms of local operations on nodes.

The result is an interlingual system as outlined in figure 19.10 for Dutch to English.

I will now give a more precise definition of isomorphy. First, a syntactic derivation tree is called well-formed if it defines at least one sentence (i.e. the rules in the derivation tree are applicable). A semantic derivation tree is called well-formed if one of the derivation trees to which it corresponds (according to the semantic component) is well-formed. Two grammars G and G′ are called isomorphic if each semantic derivation tree that is well-formed with respect to G is also well-formed with respect to G′, and vice versa. Note that isomorphy is an equivalence relation between grammars and that the definition can be extended easily to sets of more than two grammars.

The definitions imply that if a translation system as outlined in figure 19.10 is based on isomorphic grammars, we know that the analysis of a sentence in the source language yields a semantic derivation tree, the generation component will always yield a correct sentence of the target language. Translations defined in this way have the same meaning, they have the same semantic derivation tree, they have similar syntactic derivation trees, and they may have completely different surface trees. So in this framework the information that is conveyed during translation is not only the model-theoretical meaning, but also the way in which this meaning is derived. This could be called the compositionality principle of translation.

This approach avoids the earlier mentioned problems with intentional logic as an interlingua. The hardest of these problems was the "subset problem", which arises not only in a system with a logical interlingua, but also in a system with transfer on syntactic derivation trees (as in figure 19.8), if the grammars of source and target language are developed independently. In principle this problem is solved in a system based on isomorphic grammars, but it would be somewhat misleading to state it that way. A remaining problem is that in the syntactic framework we use, it is not yet possible to prove formally whether two grammars are isomorphic or not. For various kinds of grammars a formal proof is possible, but not yet for grammars with the syntactic power of M-grammars. However, even without a formal proof the approach is an important step forward.

In practice the process of grammar writing proceeds as follows. A set of compositional rules *R* is written for handling a particular phenomenon in

language L, a corresponding set of compositional rules R' is written for handling the corresponding phenomenon in language L$'$. The rules R should be complete for the expected set of input expressions, the rules R' should be complete for the corresponding set of input expressions of L$'$ (their "translations"). The most important practical difference between this and other approaches may be that here the grammars are written with translation in mind. Because of the reversibility of the grammars the rule writers can focus their attention on writing compositional (i.e., generative) rules in parallel and on the applicability of these rules to the expected inputs.

Figure 19.10 shows the global design of the systems which are being developed in the Rosetta project. This is a research project on machine translation at Philips Research Laboratories, Eindhoven. A few years of preparatory research resulted in the isomorphic grammar approach outlined here and in two experimental systems based on this approach, Rosetta1 and Rosetta2. A fairly large six-year project has started this year (1985), in which more sophisticated systems, Rosetta3 and Rosetta4, will be developed, for Dutch, English and Spanish.

The Rosetta approach is interlingual. Since interlinguality can be defined in various ways, this statement may cause misunderstandings. Therefore I will give three possible definitions of interlinguality and indicate which of them are applicable here.

1. A system is interlingual if there is an intermediate meaning representation which has the "same distance" to the sentences of the source language and the target language. Note that according to this definition even a bilingual one-direction translation system may be interlingual. This definition is clearly applicable to the Rosetta systems.

2. A system is interlingual if an interlingua is defined for a given set of languages in such a way that for each of these languages an analysis component can be defined that translates from that language into the interlingua and a generation component that does the reverse. So the combination of an analysis component for language L and a generation component for language L$'$ is a translation system from L into L$'$. This definition is also applicable to the Rosetta systems.

3. A system is interlingual if it uses an interlingua which is "universal", i.e., which can be used for expressing any meaning of any sentence in any natural language. Obviously, the Rosetta approach is not interlingual in this sense.

In the Rosetta project we aim at developing an interlingual system, according to definition 2. This is certainly more ambitious and more difficult than developing a purely bilingual system. Rosetta3 is being developed for three languages in order to find out what the price of this multilingual approach is, in comparison with the bilingual approach according to definition 1.

Concluding Remarks

In the section on Montague Grammar and machine translation I formulated three requirements on translation systems: explicit grammars of source language and target language, translation of correct sentences into correct sentences according to these grammars, and a definition of what has to be conveyed during translation. The isomorphic grammar approach satisfied the first two requirements; with respect to the third requirement a step forward has been made in comparison with using Intentional Logic as interlingua. In the Rosetta systems it is not only the model-theoretical meaning that is conveyed, but also the way in which this meaning is derived from basic meanings.

I mentioned three problems with using Intentional Logic as an interlingua. The first problem was that a meaning representation in Intentional Logic may require a more detailed meaning analysis than is needed for translation purposes, because for translation we are mainly interested in equality of meanings. This problem is solved by using semantic derivation trees as interlingual meaning representations, in which the unique names of basic meanings and meaning rules serve exactly to express the equality of meaning of basic expressions and syntactic rules, respectively. The second problem was that expressions of Intentional Logic only convey the meaning in the strict model-theoretical sense. As I pointed out, semantic derivation trees indicate in addition the way in which the meaning is derived. They may also be used to convey other information than the meaning. If two basic expressions or two syntactic rules (of the same language) have the same meaning, but differ in some other aspect which is relevant to translation, we may assign different names to the corresponding basic meanings and meaning rules. The solution of the third problem, the subset problem, has been the main motivation for the isomorphic grammar approach. If the grammars of the source and the target language are isomorphic, each interlingual expression generated by

the analysis component can be processed by the generation component.

In this paper I have illustrated the isomorphic grammar approach by means of very simple examples. This may leave you with the impression that isomorphic grammars can only define very trivial translation relations. The following remarks should indicate the potential power of the approach.

1. First and foremost it is important to notice that the rules and the basic expressions of the grammars are chosen with translation in mind.

2. Syntactic rules may perform powerful operations on syntactic trees, e.g., permutations, substitutions and deletions, as long as the conditions on M-grammars are obeyed. So the correspondence between syntactic rules of different grammars as required by the isomorphy relation does not imply similarity of the surface structures.

3. Basic expressions need not be terminals (i.e. S-trees consisting of one node), but may also be complex S-trees. This is especially useful for idiomatic expressions (e.g., *to make up one's mind*), which are primitive from a semantic point of view, but complex from a syntactic point of view. The same mechanism is used in the case where a word in one language corresponds to a complex expression in the other language, even if this complex expression would not be considered as an idiom in that language (e.g. the translation of the Spanish verb *madrugar* into the English expression *modulator*). On the other hand basic expressions may correspond to "deeper", possibly more abstract, notions than those denoted by words.

4. Corresponding basic expressions of two languages need not have the same syntactic category, under the conditions that these different categories correspond to the same semantic type. Obviously allowing such a mismatch of categories imposes conditions on the rest of the grammars which are not always easy to fulfill. In Landsbergen (1985) I dealt with a particular example of this: the translation of the English verb *to like* into the Dutch adverb *graag*.

I hope that these points make it clear that the isomorphic grammar approach is in principle quite powerful. The practical feasibility should be shown and has to a certain extent been shown already by the actual systems developed in the Rosetta project.

In conclusion, I hope to have shown that application of Montague Grammar in machine translation may yield the best results if it is applied in a "creative" way. The main influence on the Rosetta systems has been exerted by the compositionality principle. This plays an important role in Rosetta, not only by relating form and meaning of one language, but also by inspiring us to formulate a compositionality principle of translation which relates form and meaning of various languages. These principles should not be interpreted as refutable theories of language or translation (cf. Partee 1982 on the status of the compositionality principle), but as guiding principles for the construction of grammars and translation systems.

And there I would like to stop.

Discussion

Pete Whitelock: Well, I would like to ask a question. In your approach, if you have a sentence which is ambiguous in translation but non-ambiguous in the source language as far as we can tell, do you have to essentially give it two analyses so that you can get the two translations?

Landsbergen: If a sentence is ambiguous in translation, i.e., if it has more than one translation, there are two possibilities. The first one is that these translations are paraphrases, corresponding to the same meaning. In that case there is only one analysis of the SL sentence and the ambiguity arises only in the generation component. The second possibility is that these translations correspond to different meanings. In that case there must be two analyses of the SL sentence. It is not always easy to decide if for a particular phenomenon we have to create a semantic ambiguity or if it can be described as having one "encompassing" meaning. In Rosetta this decision will not only depend on what is most elegant in one language, but it will also be influenced by the other languages.

Doug Arnold: The language that the grammar defines is something rather close to the surface of the languages—it's something, I imagine, like morphologically and syntactically analyzed English, or morphologically and syntactically analyzed Dutch, and so on. That's right, isn't it? You have only one set of ...

Landsbergen: One level of representation, yes. However, during the generation process of a sentence, we start with rather abstract representations, which are gradually transformed into surface representations. But they are all S-trees, so essentially there is one level of representation.

Arnold: What is your feeling about having more levels of representation, so that in fact the 'tuning' of the grammars would be between grammars that essentially generate semantic represention of appropriate languages or, let's say, F-structures of the languages, logical forms of the languages, something like that? Do you have an argument against using other levels of representation, for instance?

Landsbergen: Well, in the first place it is the other way round. There should be arguments for having more levels. But leaving that aside: in Rosetta the syntactic rules have a clear effect on both the form and the meaning. If there are more levels between form and meaning the effect of the rules may be harder to understand. But the main problem with having more levels is the "subset problem" I discussed in my presentation. If there are more levels, the representations at the deepest level will be the result of a number of translation steps between the various levels. It is hard to characterize independently the subset of deep representations that correspond to sentences. This makes it difficult to guarantee that this subset is actually translated.

Arnold: I think that the subset problem is one of the major problems. Could I just say what the argument for having other levels is: there are more superficial differences between languages than there are non-superficial ones; so languages configure differently, let's say. So a nonconfigurational representation makes translation easier. You can phrase that within a different theory if you want, but there is that sort of intuition around. That would motivate having other levels than one.

Landsbergen: I forgot to mention another objection against having deeper levels. After going to a deeper level of analysis, information that is useful for translation may get lost. E.g., at the F-structure level of LFG there is no information about the surface order of constituents, although this may be important for choosing the most plausible interpretation with regard to scope. Of course, the idea that languages have more in common at a deeper level of analysis than at the surface is an argument in favor of having more levels. But in our approach the derivational history is such a level; our assumption is that languages have much in common at the level of derivational history.

But I interrupted you—please continue.

Arnold: My point really relates to the subset problem. Why don't you just say, for the cases where there is a failure of intersection between source language and target language ILs, that there is no translation in those cases? Why don't you adopt a more restrictive view of translation, distinguishing, say, between translation and paraphrase?

Landsbergen: There are two reasons. The first reason is a practical one. We make an interlingual system with interaction with the user during analysis, in case of ambiguities. If in such a system the analysis has been successful and has yielded an interlingual expression, one wants to be sure that the generation component provides a translation.

Arnold: Why? If what you are doing is translation why don't you ...

Landsbergen: Well, I think of the application of this system in an electronic mail environment. It is unacceptable if an analyzed message is not translated.

Arnold: No, I was pressing you for a theoretical argument.

Landsbergen: OK, that was a practical point.

Arnold: Why do you call the result of that sort of activity "translation" and not something else? If the source text and a target text don't share at least one IL representation, why do you want to claim that they are translations?

Landsbergen: The theoretical argument is that if the source text and the target text do not share an IL representation, it may still be the case that they have logically equivalent representations. So in that case they have the same meaning and may be called each other's translations, but due to fairly arbitrary differences in the two grammars, they are not recognized as such by the system.

Henry Thompson: I suspect that really the right place to get an answer to this is in Partee's work, but on a quick understanding of what you said, can you disabuse me of the notion that a Montague Grammar with constraints imposed on it to ensure parsability is any different from a context-free grammar? Is there an obvious way to characterize the difference between a Montague Grammar so restricted, particularly the S-rules that are associated with it, and something that I would think of as a context-free grammar with a rule-to-rule relationship between the syntactic rules and some compositional semantics? Is there anything that really remains of Montague in this? That is, I guess, what it comes down to.

Landsbergen: Montague's own example grammars are more or less context-free, but in Rosetta we use a

transformational extension of Montague Grammar (cf. Landsbergen 1981). Our rules are powerful, they can perform permutations, deletions etc. Indeed, our surface grammar is context-free in its weak generative capacity, but the grammar as a whole defines a non-context-free subset of this. Actually, our formalism is undergoing some changes at the moment. We are going to make a distinction between meaningful rules that contain information relevant for translation and on the other hand purely syntactic transformations. These transformations are not involved in the isomorphy relation and can be defined for each language separately.

Thompson: What does the parser then look like as a result of all this?

Landsbergen: The parser consists of two parts: the surface parser and the M-parser. The surface parser produces a set of candidate surface trees for the input sentence. The M-parser applies the analytical rules of the M-grammar to a surface tree and breaks it down into smaller parts, ultimately into basic expressions. If the M-parser is successfully applied, i.e., if the surface tree is correct, the result is a derivation tree. The surface grammar is weakly equivalent to a context-free grammar, it is similar to a recursive transition network grammar. The rules of the M-grammar are more powerful.

Thompson: Thank you.

Graeme Ritchie: Could I ask you about idioms? I'm a bit puzzled about what you said about idioms. It sounded from what you said as if, if one of the languages had a phrasal, idiomatic expression of some concept, there had to be a basic concept in the logic and a basic expression in the semantics corresponding to that which had that semantic compositional structure.

Landsbergen: No no no. Not that.

Ritchie: Well you said that idioms may have whole semantic derivation trees.

Landsbergen: I said that idioms correspond to compound basic expressions. I am sorry about all these different kinds of trees, but here we have to make clear distinction between S-trees and derivation trees. All basic expressions are S-trees, but usually they consist of one node. An idiom is a compound S-tree, consisting of more than one node. It is a basic expression from a semantic point of view, but it is a compound expression from a syntactic point of view. For example, *to lose one's temper* will be represented as an S-tree with *lose* and *temper* in it, but its meaning is not derived compositionally from these parts.

Ritchie: For the semantics that's derived from it, to do the translation the other language has to have some expression which has that as its semantics?

Landsbergen: Yes, the expression in the other language may be atomic or may be an idiomatic expression. It may also be a compound expression that one would not be inclined to call an idiomatic expression in that language. For instance, a possible translation of *to lose one's temper* into Dutch is *kwaad worden*, an idiomatic expression of Dutch, but in the translation system it has to be treated in the same way as an idiom.

Ritchie: I can understand that. I didn't see what the adjective "compound" implied with your various levels.

Karen Sparck Jones: You said quite explicitly you're not dealing with ill-formed text at the moment, fragments and things like that. Is it perfectly obvious how, when you've got around to it, in principle you would do this in this kind of approach?

Landsbergen: I did not deal with ill-formed input in my paper, but in the actual system Rosetta2 we try to deal with it. For sentences that do not fit into the system's grammar, there are several robustness measures, partially similar to those in other systems. For instance, if the surface parser is not able to make a complete parse, it will look for a "cover" of the sentence by the largest constituents it has found. It puts them together under a special node with category UG (for "Ungrammatical"). In the next phase, the M-parser, there is an analytical rule that is able to cope with a UG. At the moment this rule is very simple: it splits up the tree into its immediate subtrees. Each of the subtrees is then analyzed and translated further in the usual way. In the generation component the translated subtrees are combined again by a rule corresponding to the beforementioned analysis rule for a UG. So the net result of all this is that an incorrect sentence is split up into correct parts which are translated separately.

Nick Ostler: Do you have any experience of working practically with, say, three languages? I don't know whether it's only in the future that you are going to bring in Spanish, but it seems that you envisage a real-time interaction between linguists working together drafting these grammars, and presumably that's just about feasible when you've got two languages. If you've got three, establishing your

isomorphisms will be twice as difficult again, I suppose, and if you were to add more languages of course it would rapidly become completely infeasible.

Landsbergen: I have some experience with writing isomorphic grammars for Dutch, English, and Italian, for Rosetta1, but these grammars were small and I did that on my own, so there I did not encounter the problems you are talking about. The second version of the system, Rosetta2, has larger grammars, which have been designed for the same three languages, but they have been worked out only for Dutch and English, due to a change in our planning. We are now working with a group of linguists and the actual writing of the rules has to start yet. We will first make global isomorphic schemes for the three grammars. Then these grammars will be worked out in detail, separately. If serous problems arise in that phase, there may be feed-back to the isomorphic scheme.

Ostler: But you haven't done it very much as yet? This is your plan for the six-year project.

Landsbergen: Yes. The six-year project itself is very young. It started at the beginning of this year [1985].

Ostler: So your experience is just of doing English and Dutch. There has been the PHLIQA project (Landsbergen 1976).

Landsbergen: That was in a way the predecessor of this project.

Ostler: Did that involve multilingual or just bilingual . . .

Landsbergen: No, it was just English. PHLIQA was a question-answering system. So we have experience with building large systems, but not with building a large interlingual translation system with a group of linguists. Note that the isomorphic approach is also feasible for bilingual translation. We have chosen to work on three languages, because we are interested in interlingual applications and want to investigate to what extent the multilingual approach is feasible. One of the goals of the project is to find out what the price of this multilinguality is. I hope to report on this in a few years.

Note

1. The Rosetta project is partially sponsored by NEHEM (Nederlandse Herstructureringsmaatschappij). I would like to thank Jeroen Groenendijk, Kees van Deemter, Rene Leermakers, and Jan Odijk for their comments.

References

Dowty, D. R., R. E. Wall, and S. Peters. 1981. *Introduction to Montague Semantics*. Dordrecht: D. Reidel.

Friedman, J., and D. S. Warren. 1978. A Parsing Method for Montague Grammars. *Linguistics and Philosophy* 2, 347–372.

Godden, K. 1981. Montague Grammar and Machine Translation between Thai and English. Ph.D. dissertation, University of Kansas.

Janssen, T. M. V. 1986. Foundations and Applications of Montague Grammar, Part I. CWI Tract 19. Amsterdam: Centre for Mathematics and Computer Science.

Landsbergen, S. P. J. 1976. Syntax and Formal Semantics of English in PHLIQA1. In L. Steels, ed., *Advances in Natural Language Processing*. Antwerp: University of Antwerp.

Landsbergen, S. P. J. 1981. Adaptation of Montague Grammar to the Requirements of Parsing. In J. A. G. Groenendijk et al., eds., *Formal Methods in the Study of Language*, *Part 2*. MC Tract 136. Amsterdam: Mathematical Centre.

Landsbergen, S. P. J. 1985. Isomorphic Grammars and their Use in the Rosetta Translation System. In M. King, ed., *Machine Translation Today*. Edinburgh: Edinburgh University Press.

Montague, R. 1970a. Universal Grammar. Theoria 36, 373–398. Reprinted in Thomason, ed., 222–246.

Montague, R. 1970b. English as a Formal Language. In B. Visentini et al., eds., *Linguaggi nella società e nella technica*. Milan: Edizioni di Comunità, 189–224. Reprinted in Thomason, ed., 108–221.

Montague, R. 1973. The Proper Treatment of Quantification in Ordinary English. In J. Hintikka et al., eds., *Approaches to Natural Language*. Dordrecht: D. Reidel, 221–242. Reprinted in Thomason, ed., 247–270.

Nishida, T. and S. Doshita. 1982. An English-Japanese Machine Translation System Based on Formal Semantics of Natural Language. In J. Horecky, ed., *COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics*. Amsterdam: North-Holland, 277–282.

Partee, B. H. 1976. Some Transformational Extensions of Montague Grammar. In B. H. Partee, ed., *Montague Grammar*. New York: Academic Press, 51–76.

Partee, B. H. 1982. Compositionality. In F. Landman and F. Veltman, eds., *Varieties of Formal Semantics* (*Proceedings of the 4th Amsterdam Colloquium*). Dordrecht: Foris, 281–312.

Thomason, R. H., ed. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. New Haven: Yale University Press.

20
Dialogue Translation vs. Text Translation—Interpretation Based Approach

Jun-ichi Tsujii and Makoto Nagao

Introduction

Although we had been engaged in developing an MT system of texts for several years (Mu project [Nagao85, Nagao86]), we were puzzled when we examined the data of dialogue translation gathered by the research group of ATR, which is a newly established research organization for translation of telephone dialogues and is now gathering dialogue translation data in various hypothetical situations.

The sample translations gathered by the ATR research group looked very difficult for machines, but we rarely found syntactic structures which make textual translation so difficult, such as long noun phrases or clauses, complicated conjuncted phrases, etc. ([Tsujii84] [Tsujii87]). On the other hand, most of the translations of dialogues between Japanese and English, which were produced by professional human interpreters, did not preserve syntactic structures of their original sentences at all. They were completely paraphrased in the target language and seemed very hard to be produced by conventional techniques developed for textual MT systems.

Both translations, dialogue and textual translations, are difficult, but their difficulties are very different from each other.

We discuss in this paper the differences of dialogue translation systems and textual translation systems. Because we do not know the difficulties of recognition of spoken utterances, we will avoid the discussion about the difficulties of interfacing the speech recognition part and the linguistic processing part, which we will certainly encounter in spoken dialogue translation systems. The dialogue translation in this paper is restricted to the translation of dialogues through keyboards, on which ATR is now concentrated.

The differences of these two translation systems mainly come from the fact that dialogues of certain types are more goal-oriented than ordinary texts. We will argue that the goal orientedness of dialogues makes dialogue translation systems more feasible than textual translation systems, though they are usually considered much harder.

Differences of Environments

In the current states of the art in machine translation, most researchers may agree that we cannot expect an ideal FAMT system which can translate any linguistic materials in any subject domains. So, at present, what should be discussed about MT systems have to be *engineering problems*.

We should discuss problems from engineering points of view. That is, we should discuss, first of all, what types of systems or system organizations are economically and technically feasible in what situations of actual translation, and what sorts of human aids can be expected in real application environments.

The important consideration is how to design feasible MT systems which can be used in actual, rather specific, translation *environments*. Different application environments require different technologies. Therefore, the questions we would like to pose in this paper are:

- Which is more feasible in *actual application environments*, dialogue translation systems or textual translation systems?
- Can we design a feasible dialogue translation system just by extending or modifying current MT technologies developed exclusively for textual translation?

Our answer to the first question, though it might sound strange, is that dialogue translation systems of certain types are more feasible than textual translation systems which are currently developed and commercially available.

It might be the case that we imagine dialogue translation is easier, because we have been engaged in developing a textual translation system and have recognized many, not only difficult but also nasty and dirty problems in textual translation systems [Nakamura86].

But not only because of that, we believe dialogue translation systems are more feasible, mainly because of the basic differences of environments where these two types of systems will be used.

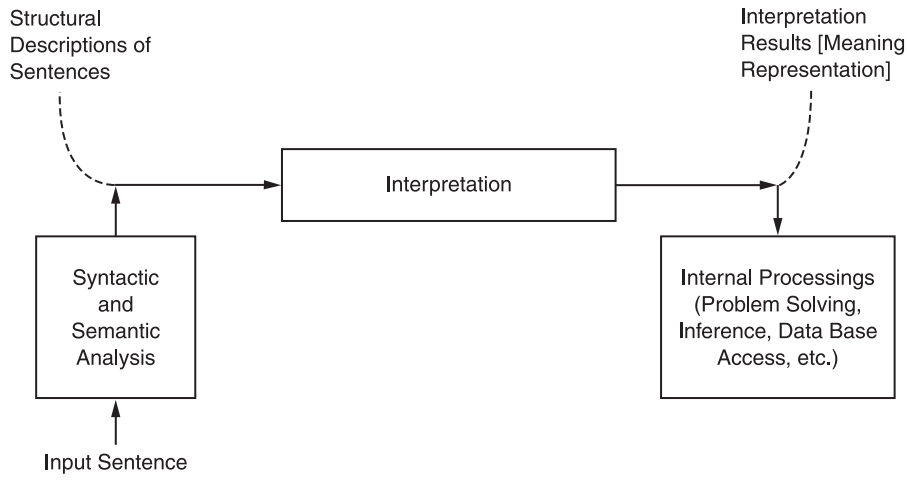


Figure 20.1

We can summarize the differences of environments in which these two types of systems might be used as follows.

• *Clear Definition of Information:* In certain types of dialogue translations, we can define rather clearly what information should be transmitted from source sentences to target translations, while we generally cannot in textual translation. By certain types of dialogues, we mean here the dialogues such as dialogues for hotel reservation and conference registration which are currently picked up by the ATR research group, dialogues between patients and doctors tried by the CMU group ([Tomita86]), etc.

• *Active Participation of Speakers and Hearers:* In most application environments of textual translation systems, they are supposed to be used by *professional translators*. We cannot have the writers of texts at the time of translation, the persons who prepare texts and really want to communicate something through the texts. The actual readers of translated texts are not available, either, at the time of the translation, who really want to get messages or *information* encoded in the texts.

On the contrary, in dialogue translation, we have both the speakers (the senders of messages) and the hearers (the receivers of messages) at the time of translating messages.

These two differences make, we claim, dialogue translation systems more feasible in actual translation environments, if they are properly designed for taking these advantages.

Our answer to the second question is directly derived from the above discussion. That is, in order to take the advantages of dialogue translation, the system organizations should be different from those for textual translation. Mere extension of current MT technologies for textual translation will not result in high quality dialogue translation systems by which *ordinary people* can communicate with each other.

We will discuss what implications the basic differences of environments have in the design of dialogue translation systems, and substantiate the conclusion that *if they are properly designed*, *certain types of dialogue translation systems are more feasible*, technically at least, than the text translation systems which are currently available.

What Should be Translated?

Figure 20.1 shows a simplified framework of application systems of natural language understanding (NLU) other than MT systems. In this framework, *understanding of a sentence* is regarded as a process of transformation from an input sentence, a linear sequence of words, into the so-called *meaning representation of the sentence*.

Meaning representation in this framework is the input to certain *internal processings* such as deductive inferences, problem solvings in certain restricted domains, data base accesses, etc., which are actually implemented as computer programs to carry out certain specific internal tasks.

Meanings of input sentences are defined in this framework, relative to the internal tasks that the sys-

tems are expected to perform. In other words, what kinds of information should be extracted from sentences are predefined, depending on the aims of the internal processings of the systems. Understanding is regarded as an extraction process of information relevant to specific internal tasks.

However, the internal task or the aim of translation is

to re-express by using sentences of target languages the information of all aspects conveyed by sentences of source languages, with as least distortion as possible ([Tsujii86]).

The internal task of MT, by itself, does not define what information should be extracted from input texts. It is commonly recognized by linguists that all different surface linguistic expressions convey different *meaning*. MT systems, at least textual translation systems, have to extract all the factors relevant to the determination of surface linguistic expressions.

Most of the difficulties peculiar to MT, such as the selection of appropriate target lexical items or expressions, etc., come from the fact that we cannot define in MT what aspects of information in source sentences are relevant to the determination of target expressions and should be extracted from source sentences. In general, we cannot establish a representational framework which is language universal and by which understanding results are represented.

As a consequence, most of the current systems use certain linguistic levels of structural descriptions of source sentences, such as *deep case structures* in the Mu project, in order to calculate appropriate target descriptions. Because the structures are far from representing *understanding results* and reflect the linguistic structures of source sentences, their translation results are inherently *structure bound*.

On the other hand, in certain types of dialogues, we can define by the purpose of dialogues what is essential or important information conveyed by utterances and should be transmitted to their translations. Here, we do not discuss the systems which are capable of translating arbitrary dialogues like chatterings among house wives without any purpose, but the systems which translate dialogues of certain restricted domains as already mentioned, such as dialogues for hotel reservation, conference registration, etc. In such dialogues, we can define *important information* by referring to the aim of the dialogues.

Such *important information* should be extracted from the input and properly transmitted to the target. So, the framework for dialogue translation becomes similar to that of the other applications of NLU illustrated in figure 20.1. We can introduce a layer of explicit understanding to MT systems, to which *important information* of utterances are related and so, in which results of *understanding* can be represented in a language-independent (but task-dependent) way ([Tsujii87]).

Some parts of utterances which convey *information important* for the purposes of the dialogues are related to this layer and *interpreted*. Because information is expressed language-independently in this layer, we can expect *less structure bound translation results* for the parts of utterances. On the other hand, the other parts which do not convey important information need not be related to this explicit understanding layer. They would be translated by conventional MT technologies.

Let me show you a simple example from hotel reservation dialogues, which actually appears in the experiments conducted at ATR.

[EX 1]

[Japanese] hoteru (*hotel*) -wa, tomodachi (*friends*)-to Disuko (*discotheque*)-ni ikitai (*to want to go to*)-node, Roppongi (*Roppongi—the name of the place in Tokyo*)-no chikaku (*to be near*)-ga iino (*to be good*)-desuga?

[Structure Bound English Translation] As for hotel, because [I] would like to go to Discotheque with friends, to be near to Roppongi is good.

[English Translation] Because I would like to go to discotheque with friends, I prefer to stay at a hotel near to Roppongi.

In this example, we can divide the utterance into two. One is the part which contains important information for hotel reservation, and the other is the part which does not. Because the location of the hotel at which the client wants to stay is important for the task of hotel reservation, the underlined part of the utterance is important and should be translated as properly as possible.

The other part of the utterance, which gives the reason why the client wants to stay at a hotel in a specific region of Tokyo (Roppongi), is less important. Our contention is that these two parts of the utterance should be treated differently in dialogue translation systems.

Note that the English translation given above has a *deep case structure* completely different from that

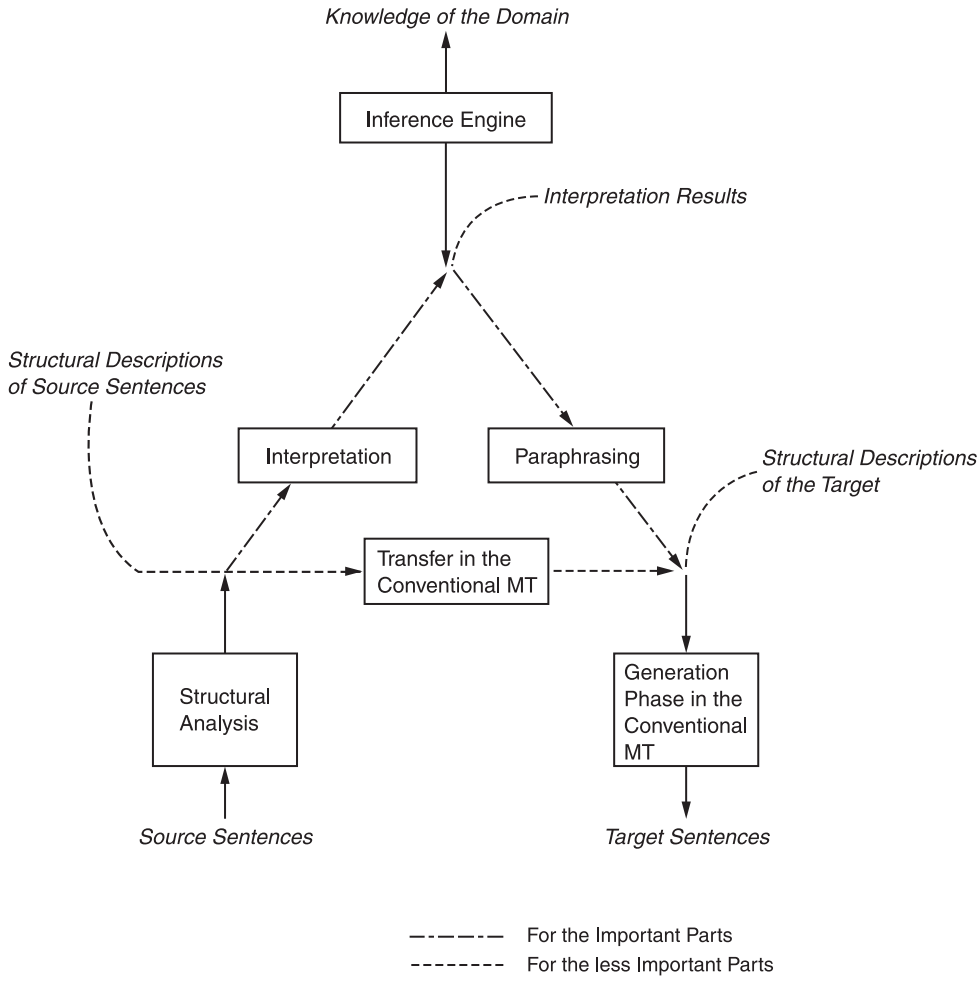


Figure 20.2

of the source sentence. The translation contains the verbs *to prefer* and *to stay* whose corresponding Japanese verbs do not appear in the source sentence.

Architecture of Dialogue Translation Systems

Figure 20.2 shows a schematic view of a system which translates dialogues in a certain restricted domain. The translation system knows in advance what kinds of *information* or *concepts* are important for the natural flow of dialogues in that specific task domain, and also knows a set of surface linguistic expressions which may convey such *important information*.

By using these kinds of knowledge, the system should be able to distinguish the parts which convey important informational contents, extract them and relate them to the representations of the explicit understanding layer.

It is certainly difficult to capture the important parts of utterances and *understand* them, but if we confine ourselves to a certain restricted task domain, it is much easier than story understandings in general, which AI researchers have been interested in.

Furthermore, it is easier than developing *intelligent dialogue systems* which make conversations with human users in restricted task domains, for example, to make appropriate hotel reservations. Although those intelligent systems should be able to understand fully the user's utterances, a dialogue *translation* system need not. The hearer, the receiver of the translated messages may understand the speaker's intention. A translation system is only required to provide *information* sufficient for his understanding. It is desirable but not inevitable for a dialogue translation system to have the ability of recognizing the speaker's plan.

A translation system which extracts *important information* from source utterances and re-expresses it in the target language can produce less structure bound translations. It can reduce varieties of surface expressions to a single *meaning representation*, if they convey essentially the same information, the *same* from the view point of the purposes of dialogue. For example, the following Japanese expressions, which have quite different (deep case) structures, may be reduced to a single representation and re-expressed by English expressions. The English expressions will be chosen independently of Japanese structures but only by considering English contexts where the expressions are located.

[EX 2]

[Japanese] Roppongi-no chikaku (*to be near*)

-no hoteru (*hotel*) -ga ii (*to be good*) wa.

→[Structure bound translation] A hotel near to Roppongi is good.

[Japanese] Roppongi-atari (*around*) -no hoteru (*hotel*) -wo onegaishimasu (*please*).

→[Structure bound translation] A hotel around Roppongi, please.

[Japanese] hoteru (*hotel*) -ha roppongi-no chikaku (*to be near*) -ga iiuodesuga (*to be good*)

→[Structure bound translation] As for hotel, to be near to Roppongi is good.

[Japanese] tsugou-ga-iino (*to be convenient*) -ha roppongi-ni chikai (*to be near*) hoteru (*hotel*) desuga.

→[Structure bound translation] What is convenient is a hotel near to Roppongi.

As an extreme, we can imagine a system which produces fluent translations only for important parts of utterances but awkward ones for the other parts.

[EX 3]

[Because (to Go Discotheque) Friends] I prefer to stay at a hotel near to Roppongi.

Note that a dialogue translation system need not understand utterances completely, and so, it need not understand why the clause 'tomodachi-to disuko-ni ikitai' (I would like to go to discotheque with friends) can be the reason for staying at a hotel near to Roppongi. To understand this, a system has to have a lot of real world knowledge which is not so closely related with hotel reservation tasks, such as

1. Roppongi is a special region in Tokyo where many discotheques exist.
2. In order to go to some place, it is preferable to stay at a hotel near to the place.
3. If something is preferable, the client tends to....

etc.

A system which converses intelligently with human to make hotel reservations should have such knowledge and abilities of using it. However, a dialogue translation system has only to provide information to the human participants who organize conversation intelligently.

Active Participation of Speakers and Hearers

What should be *understood* from texts is highly dependent on the intentions of actual writers and readers of texts, but neither of them is available at the time of translation in textual translation.

The same texts would be read by different readers with different intentions who would like to get different sorts of information from the translated texts.

Readers of translated texts are often irritated because they cannot get necessary information for them. We found that translated texts are irritating not only because translations are awkward, but also because original texts themselves do not contain *information* which actual readers would like to get. Furthermore, evaluating translations produced by MT systems is difficult, because the evaluation highly depends on both what readers want to know and what source texts really contain. MT systems cannot produce good translations from bad source texts.

However, the *environments* of dialogue translation systems, in which both actual writers and readers are available at the time of translation, are much better than textual translation. The readers can ask questions directly to the writers in order to get necessary information, when they cannot get it from the translated messages or when they cannot understand the translations.

Furthermore, the translation system can also pose questions to the writers (senders of messages) to clarify their intentions. We can expect an intelligent translation system to play a role of a coordinator of conversations by keeping track of exchanges of *important information* between dialogue participants (see figure 20.2).

[EX 4]

[English participant] In which region do you want to stay in Tokyo?

[Japanese participant] disuko-ni ikitai. (I would like to go to discotheque.)

[System's Question to the participant] shitsumon-ha anata-no kibou-suru hoteru-no basho desu? (The question is 'in which region do you want to stay in Tokyo?' Would you specify the place which you prefer to stay?)

[Japanese participant] disuko-ni chikai hoteru-desu. (A hotel near to a discotheque.)

[Translated reply to the English participant] I prefer to stay a hotel near to a discotheque.

Note also that what is important in dialogue translation is the exchange of information through translation but not translated texts obtained as the result. Translations are satisfactory when the participants achieve their goals, even if they are awkward. On the contrary, in textual translation, translated texts themselves are important and they should be natural and clear enough in all aspects, because different readers with different intentions will read them and be interested in different aspects of informational contents of same texts.

Conclusions

We discussed in this paper the differences of dialogue translation systems and textual translation systems. Especially, we emphasized the differences of environments where these two types of systems will be used, and discussed what implications the differences have in the design of feasible dialogue translation systems. The main differences are:

- Clear Definition of Information in Dialogue Translation
- Active Participation of Speakers and Hearers in Dialogue Translation

We argued that, if they are properly designed to take these advantages of dialogue translation, dialogue translation systems can be more feasible than textual translation systems. Especially, we proposed a new approach to MT, called *interpretation based approach*, in which an explicit layer of understanding is introduced and parts of utterances conveying *important information* are *interpreted* by being related to this layer.

Though the approach produces *less structure bound* translations through *understanding and paraphrasing*, it is different from the conventional pivot or interlingual approach which claims their understanding results can be represented in the forms which are independent on both individual languages and tasks. The understanding layer in the proposed approach, on the other hand, is *language universal* but highly dependent on specific tasks of dialogues. In the proposed approach, we have to design an internal meaning representation specific to the domain of the dialogues.

The following are important in order to develop a feasible dialogue translation system based on the interpretation based approach.

- Integration of different layers of descriptions: We have to devise technologies for integrating the descriptions of the understanding layer and the conventional structural descriptions of source sentences to produce translations, because single utterances generally consist of the parts which convey *important information* and those which do not. The idea of *safety net* should be re-considered in this new context.
- Flexible interaction during translation: Traditional post- and pre-editings by human are not the best way to take the advantage of the availability of speakers and hearers in dialogue translation. We have to design much flexible interaction modes including *clarification dialogues* between the system and the dialogue participants.
- Management of dialogue structures: In order to find *important information*, a system should have the ability of managing the dialogue structures. It should be able to utilize various kinds of knowledge such as knowledge about surface clue expressions, task dependent knowledge, discourse structures, etc. to recover the structures of on-going dialogue. Especially, a translation system as a coordinator of conversations has to keep track of *important information exchanges* through sequences of utterances.

Acknowledgments

The authors are grateful to the members of the study group on *Spoken Dialogue* at ATR. The discussions with them were very helpful to improve the paper.

References

[Nagao85] Nagao, M., J. Tsujii, and J. Nakamura. 1985. The Japanese Government Project for Machine Translation, *Computational Linguistics*, 11, no. 2–3.

[Nagao86] Nagao, M., J. Tsujii, and J. Nakamura. 1986. The Transfer Phase of the Mu Machine Translation System, *Proc. of COLING 86*.

[Nakamura86] Nakamura, J., J. Tsujii, and M. Nagao. 1986. Solutions for Problems of MT Parser, *Proc. of COLING 86*.

[Tomita86] Tomita, M., and J. Carbonell. 1986. Another Stride Towards Knowledge-based Machine Translation, *Proc. of COLING 86*.

[Tsujii84] Tsujii, J., J. Nakamura, and M. Nagao. 1986. Analysis Grammar of Japanese in the Mu-project, *Proc. of COLING 86*.

[Tsujii86] Tsujii, J. 1986. Future Directions of Machine Translation, *Proc. of COLING 86*.

[Tsujii87] Tsujii, J. 1987. What Is PIVOT? *Proc. of MT Summit*, Hakone, Japan.

This page intentionally left blank

21
Translation by Structural Correspondences

Ronald M. Kaplan, Klaus Netter, Jürgen Wedekind, and
Annie Zaenen

Introduction

In this paper we sketch an approach to machine translation that offers several advantages compared to many of the other strategies currently being pursued. We define the relationship between the linguistic structures of the source and target languages in terms of a set of correspondence functions instead of providing derivational or procedural techniques for converting source into target. This approach permits the mapping between source and target to depend on information from various levels of linguistic abstraction while still preserving the modularity of linguistic components and of source and target grammars and lexicons. Our conceptual framework depends on notions of structure, structural description, and structural correspondence. In the following sections we outline these basic notions and show how they can be used to deal with certain interesting translation problems in a simple and straightforward way. In its emphasis on description-based techniques, our approach shares some fundamental features with the one proposed by Kay (1984), but we use an explicit projection mechanism to separate out and organize the intra- and inter-language components.

Most existing translation systems are either transfer-based or interlingua-based. Transfer-based systems usually specify a single level of representation or abstraction at which transfer is supposed to take place. A source string is analyzed into a structure at that level of representation, a transfer program then converts this into a target structure at the same level, and the target string is then generated from this structure. Interlingua-based systems on the other hand require that a source string has to be analyzed into a structure that is identical to a structure from which a target string has to be generated.

Without further constraints, each of these approaches could in principle be successful. An interlingual representation could be devised, for example, to contain whatever information is needed to make all the appropriate distinctions for all the sentences in all the languages under consideration. Similarly, a transfer structure could be arbitrarily configured to allow for the contrastive analysis of any two particular languages. It seems unlikely that systems based on such an undisciplined arrangement of information will ever succeed in practice. Indeed, most translation researchers have based their systems on representations that have some more general and independent motivation. The levels of traditional linguistic analysis (phonology, morphology, syntax, semantics, discourse, etc.) are attractive because they provide structures with well-defined and coherent properties, but a single one of these levels does not contain all the information needed for adequate translation. The D-structure level of Government-Binding theory, for example, contains information about the predicate-argument relations of a clause but says nothing about the surface constituent order that is necessary to accurately distinguish between old and new information or topic and comment. As another example, the functional structures of Lexical-Functional Grammar do not contain the ordering information necessary to determine the scope of quantifiers or other operators.

Our proposal, as it is set forth below, allows us to state simultaneous correspondences between several levels of source-target representations, and thus is neither interlingual nor transfer-based. We can achieve modularity of linguistic specifications, by not requiring conceptually different kinds of linguistic information to be combined into a single structure. Yet that diverse information is still accessible to determine the set of target strings that adequately translate a source string. We also achieve modularity of a more basic sort: our correspondence mechanism permits contrastive transfer rules that depend on but do not duplicate the specifications of independently motivated grammars of the source and target languages (Isabelle and Macklovitch, 1986; Netter and Wedekind, 1986).

A General Architecture for Linguistic Descriptions

Our approach uses the equality- and description-based mechanisms of Lexical-Functional Grammar. As introduced by Kaplan and Bresnan (1982), lexical-functional grammar assigns to every sentence two levels of syntactic representation, a constituent structure (c-structure) and a functional structure (f-structure). These structures are of different formal types—the c-structure is a phrase-structure tree while the f-structure is a hierarchical finite function—and they characterize different aspects of the information carried by the sentence. The c-structure represents the ordered arrangement of words and phrases in the sentence while the f-structure explicitly marks its grammatical functions (subject, object, etc.). For each type of structure there is a special notation or description-language in which the properties of desirable instances of that type can be specified. Constituent structures are described by standard context-free rule notation (augmented with a variety of abbreviatory devices that do not change its generative power), while f-structures are described by Boolean combinations of function-argument equalities stated over variables that denote the structures of interest. Kaplan and Bresnan assumed a correspondence function mapping between the nodes in the c-structure of a sentence and the units of its f-structure, and used that piecewise function to produce a description of the f-structure (in its equational language) by virtue of the mother-daughter, order, and category relations of the c-structure.

The formal picture developed by Kaplan and Bresnan, as clarified in Kaplan (1987), is illustrated in the following structures for sentence (1) (figure 21.1).

The c-structure appears on the left, the f-structure on the right. The c-structure-to-f-structure correspondence, ϕ, is shown by the linking lines. The correspondence ϕ is a many-to-one function taking the S, VP, and V nodes all into the same outermost unit of the f-structure, f_1.

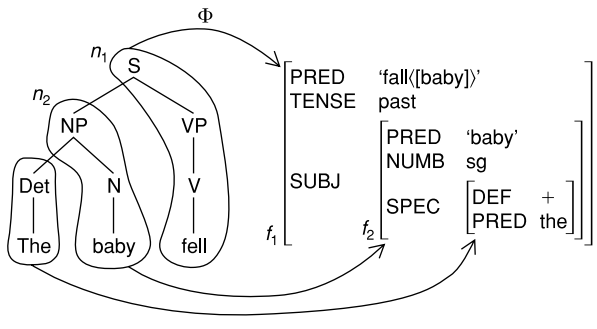


Figure 21.1

The node-configuration at the top of the tree satisfies the statement S → NP VP in the context-free description language for the c-structure. As suggested by Kaplan (1987), this is a simple way of defining a collection of more specific properties of the tree, such as the fact that the S node (labeled n_1) is the mother of the NP node (n_2). These facts could also be written in equational form as $M(n_2) = n_1$, where M denotes the function that takes a tree-node into its mother. Similarly, the outermost f-structure satisfies the assertions $(f_1 \text{ TENSE}) = \text{past}$, $(f_1 \text{ SUBJ}) = f_2$, and $(f_2 \text{ NUMB}) = \text{sg}$ in the f-structure description language. Given the illustrated correspondence, we also know that $f_1 = \phi(n_1)$ and $f_2 = \phi(n_2)$. Taking all these propositions together, we can infer first that $(\phi(n_1) \text{ SUBJ}) = \phi(n_2)$ and then that $(\phi(M(n_2)) \text{ SUBJ}) = \phi(n_2)$. This equation identifies the subject in the f-structure in terms of the mother-daughter relation in the tree.

In LFG the f-structure assigned to a sentence is the smallest one that satisfies the conjunction of equations in its functional description. The functional description is determined from the trees that the c-structure grammar provides for the string by a simple matching process. A given tree is analyzed with respect to the c-structure rules to identify particular nodes of interest. Equations about the f-structure corresponding to those nodes (via ϕ) are then derived by substituting those nodes into equation-patterns or schemata. Thus, still following Kaplan (1987), if * appears in a schema to stand for the node matching a given rule-category, the functional description will include an equation containing that node (or an expression such as n_2 that designates it) instead of *. The equation $(\phi(M(n_2)) \text{ SUBJ}) = \phi(n_2)$ that we inferred above also results from instantiating the schema $(\phi(M(^*)) \text{ SUBJ}) = \phi(^*)$ annotated to the NP element of the S rule in (2a) when that rule-element is matched against the tree in (1b). Kaplan observes that the ↑ and ↓ metavariables in the Kaplan/Bresnan formulation of LFG are simply convenient abbreviations for the complex expressions $\phi(M(^*))$ and $\phi(^*)$, respectively, thus explicating the traditional, more palatable formulation in (2b).

(2) (a) S → NP VP
$(\phi(M(^*)) \text{ SUBJ}) = \phi(^*) \quad \phi(M(^*)) = \phi(^*)$

(b) S ↑ NP VP
$(\uparrow \text{ SUBJ}) = \downarrow \quad \uparrow = \downarrow$

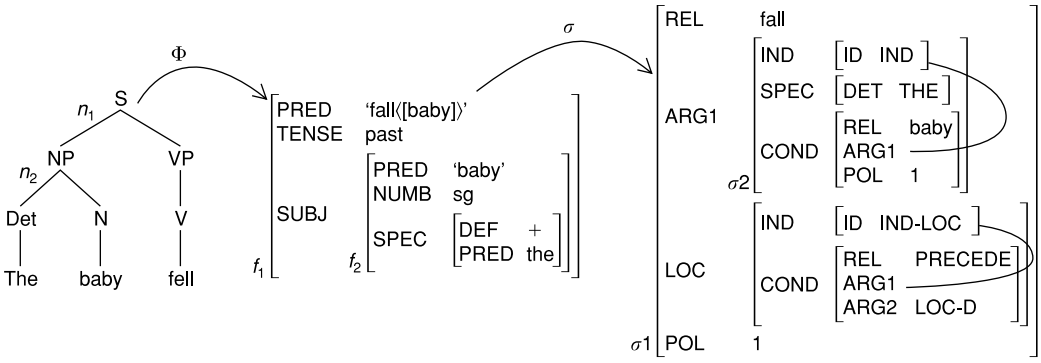


Figure 21.2

This basic conception of descriptions and correspondences has been extended in several ways. First, this framework has been generalized to additional kinds of structures that represent other subsystems of linguistic information (Kaplan, 1987; Halvorsen, 1988). These structures can be related by new correspondences that permit appropriate descriptions of more abstract structures to be produced. Halvorsen and Kaplan (1988), for example, discuss a level of semantic structure that encodes predicate-argument relations and quantifier scope, information that does not enter into the kinds of syntactic generalizations that the f-structure supports. They point out how the semantic structure can be set in correspondence with both c-structure and f-structure units by means of related mappings σ and σ'. Kaplan (1987) raises the possibility of further distinct structures and correspondences to represent anaphoric dependencies, discourse properties of sentences, and other *projections* of the same string.

Second, Kaplan (1988) and Halvorsen and Kaplan (1988) discuss other methods for deriving the descriptions necessary to determine these abstract structures. The arrangement outlined above, in which the description of one kind of structure (the f-structure) is derived by analyzing or matching against another one, is an example of what is called *description-by-analysis*. The semantic interpretation mechanisms proposed by Halvorsen (1983) and Reyle (1988) are other examples of this descriptive technique. In this method the grammar provides general patterns to compare against a given structure and these are then instantiated if the analysis is satisfactory. One consequence of this approach is that the structure in the range of the correspondence, the one whose description is being developed, can only have properties that are derived from information explicitly identified in the domain structure (see figure 21.2).

Another description mechanism is possible when three or more structures are related through correspondences. Suppose the c-structure and f-structure are related by ϕ as in (2a) and that the function σ then maps the f-structure units into corresponding units of semantic structure of the sort suggested by Fenstad et al. (1987). The formal arrangement is shown in figure 21.2. This configuration of cascaded correspondences opens up a new descriptive possibility. If σ and ϕ are both structural correspondences, then so is their composition $\sigma \circ \phi$. Thus, even though the units of the semantic structure correspond directly only to the units of the f-structure and have no immediate connection to the nodes of the c-structure, a semantic description can be formulated in terms of c-structure relations. The expression $\sigma(\phi(M(*)))$ can appear on a c-structure rule-element to designate the semantic-structure unit corresponding to the f-structure that corresponds to the mother of the node that matches that rule-element. Since projections are monadic functions, we can remove the uninformative parentheses and write $(\sigma\phi M^* \text{ ARG1}) = \sigma(\phi M^* \text{ SUBJ})$ or, using the $\uparrow$ metavariable, $(\sigma\uparrow \text{ ARG1}) = \sigma(\uparrow \text{ SUBJ})$. Schemata such as this can be freely mixed with LFG's standard functional specifications in lexical entries and c-structure rules. For example, the lexical entry for *fall* might be given as follows:

(3) *fall* V $(\uparrow \text{PRED}) = \text{'fall'}$
$(\sigma\uparrow \text{REL}) = \text{fall}$
$(\sigma\uparrow \text{ARG1}) = \sigma(\uparrow \text{SUBJ})$

Descriptions formulated by composing separate correspondences have a surprising characteristic: they allow the final range structure (e.g., the semantic

structure) to have properties that cannot be inferred from any information present in the intermediate (f-) structure. But those properties can obtain only if the intermediate structure is derived from an initial (c-) structure with certain features. For example, Kaplan and Maxwell (1988a) exploit this capability to describe semantic structures for coordinate constructions which necessarily contain the logical conjunction appropriate to the string even though there is no reasonable place for that conjunction to be marked in the f-structure. In sum, this method of description, which has been called *codescription*, permits information from a variety of different levels to constrain a particular structure, even though there are no direct correspondences linking them together. It provides for modularity of basic relationships while allowing certain necessary restrictions to have their influence.

The descriptive architecture of LFG as extended by Kaplan and Halvorsen provides for multiple levels of structure to be related by separate correspondences, and these correspondences allow descriptions of the various structures to be constructed, either by analysis or composition, from the properties of other structures. Earlier researchers have applied these mechanisms to the linguistic structures for sentences in a single language. In this paper, we extend this system one step further: we introduce correspondences between structures for sentences in different languages that stand in a translation relation to one another. The description of the target language structures are derived via analysis and codescription from the source language structures, by virtue of additional annotations in c-structure rules and lexical entries. Those descriptions are solved to find satisfying solutions, and these solutions are then the input to the target generation process.

In the two language arrangements sketched below, we introduce the τ correspondence to map between the f-structure units of the source language and the f-structure units of the target language. The σ correspondence maps from the f-structure of each language to its own corresponding semantic structure, and a second transfer correspondence τ' relates those structures (figure 21.3).

This arrangement allows us to describe the target f-structure by composing ϕ and τ to form expressions such as $\tau(\phi M^*\ \text{COMP}) = (\tau\phi M^*\ \text{XCOMP})$ or simply $\tau(\uparrow \text{COMP}) = (\tau\uparrow \text{XCOMP})$. This maps a COMP in the source f-structure into an XCOMP in the target f-structure. The relations asserted by this equation are depicted in the following source-target diagram (figure 21.4).

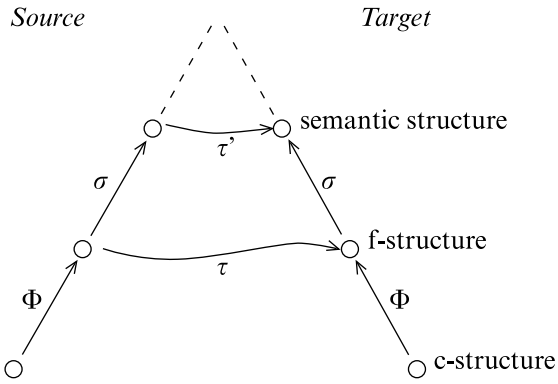


Figure 21.3

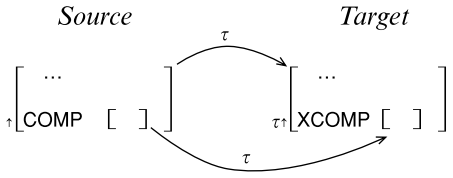


Figure 21.4

As another example, the equation $\tau(\sigma\uparrow \text{ARG1}) = (\sigma\tau\ \text{ARG1})$ identifies the first arguments in the source and target semantic structures. The equation $\tau'\sigma(\uparrow \text{SUBJ}) = \sigma(\tau\uparrow \text{TOPIC})$ imposes the constraint that the semantics of the source SUBJ will translate via τ' into the semantics of the target TOPIC but gives no further information about what those semantic structures actually contain.

Our general correspondence architecture thus applies naturally to the problem of translation. But there are constraints on correspondences specific to translation that this general architecture does not address. For instance, the description of the target-language structures derived from the source-language is incomplete. The target structures may and usually will have grammatical and semantic features that are not determined by the source. It makes little sense, for example, to include information about grammatical gender in the transfer process if this feature is exhaustively determined by the grammar of the target language. We can formalize the relation between the information contained in the transfer component and an adequate translation of the source sentence into a target sentence as follows: for a target sentence to be an adequate translation of a given source sentence, it must be the case that a minimal structure assigned to that sentence by the target grammar is subsumed by a minimal solution to the transfer description. One

desirable consequence of this formalization is that it permits two distinct target strings for a source string whose meaning in the absence of other information is vague but not ambiguous.

Thus this conceptual and notational framework provides a powerful and flexible system for imposing constraints on the form of a target sentence by relating them to information that appears at different levels of source-language abstraction. This apparatus allows us to avoid many of the problems encountered by more derivational, transformational or procedural models of transfer. We will illustrate our proposal with examples that have posed challenges for some other approaches.

Examples

Changes in grammatical function. Some quite trivial changes in structure occur when the source and the target predicate differ in the grammatical functions that they subcategorize for. We will illustrate this with an example in which a German transitive verb is translated with an intransitive verb taking an oblique complement in French:

(6) (a) Der Student beantwortet die Frage.
(b) L'étudiant répond à la question.

We treat the oblique preposition as a PRED that itself takes an object. Ignoring information about tense, the lexical entry for *beantworten* in the German lexicon looks as follows:

(7) *beantworten* V
(↑ PRED) = 'beantworten ⟨(↑ SUBJ)(↑ OBJ)⟩'

while the transfer lexicon for *beantworten* contains the following mapping specifications:

(8) (τ↑ PRED FN) = répondre
(τ↑ SUBJ) = τ(↑ SUBJ)
(τ↑ AOBJ OBJ) = τ(↑ OBJ)

We use the special attribute FN to designate the function-name in semantic forms such as 'beantworten ⟨(↑ SUBJ)(↑ OBJ)⟩'. In this transfer equation it identifies *répondre* as the corresponding French predicate. This specification controls lexical selection in the target, for example, selecting the following French lexical entry to be used in the translation:

(9) *répondre* V
(↑ PRED) = 'répondre ⟨(↑ SUBJ)(↑ AOBJ)⟩'

$$f_{58}\begin{bmatrix}\text{PRED} & \text{'beantworten⟨[Student], [Frage]⟩'}\\ \text{TENSE} & \text{present}\\ \text{SUBJ} & f_{21}\begin{bmatrix}\text{PRED} & \text{'Student'}\\ \text{NUMB} & \text{sg}\\ \text{GEND} & \text{masc}\\ \text{SPEC} & f_{81}\begin{bmatrix}\text{DEF} & +\\ \text{PRED} & \text{der}\end{bmatrix}\end{bmatrix}\\ \text{OBJ} & f_{56}\begin{bmatrix}\text{PRED} & \text{'Frage'}\\ \text{NUMB} & \text{sg}\\ \text{GEND} & \text{fem}\\ \text{SPEC} & f_{82}\begin{bmatrix}\text{DEF} & +\\ \text{PRED} & \text{die}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

$$\tau_{58}\begin{bmatrix}\text{PRED} & \text{'répondre⟨[étudiant], [à]⟩'}\\ \text{TENSE} & \text{present}\\ \text{SUBJ} & \tau_{21}\begin{bmatrix}\text{PRED} & \text{'étudiant'}\\ \text{GEND} & \text{MASC}\\ \text{NUMB} & \text{sg}\\ \text{SPEC} & \tau_{84}\begin{bmatrix}\text{DEF} & +\\ \text{PRED} & \text{le}\end{bmatrix}\end{bmatrix}\\ \text{AOBJ} & \tau_{83}\begin{bmatrix}\text{PRED} & \text{'à⟨[question]⟩'}\\ \text{PCASE} & \text{AOBJ}\\ \text{OBJ} & \tau_{56}\begin{bmatrix}\text{PRED} & \text{'question'}\\ \text{GEND} & \text{FEM}\\ \text{NUMB} & \text{sg}\\ \text{SPEC} & \tau_{85}\begin{bmatrix}\text{DEF} & +\\ \text{PRED} & \text{la}\end{bmatrix}\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

Figure 21.5

With these entries and the appropriate but trivial entries for *der Student* and *die Frage* we get the following f-structure in the source language and associated f-structure in the target language for the sentence in figure 21.5.

The second structure is the f-structure the grammar of French assigns to the sentence in (6b). This f-structure is the input for the generation process. Other examples of this kind are pairs like *like* and *plaire* and *help* and *helfen*.

In the previous example the effects of the change in grammatical function between the source and the target language are purely local. In other cases there is a non-local dependency between the subcategorizing verb and a dislocated phrase. This is illustrated by the relative clause in (11):

(11) (a) ... der Brief, den der Student zu beantworten scheint.
(b) ... la lettre, à laquelle l'étudiant semble répondre.
... the letter, that the student seemed to answer.

likely A

(↑ PRED) = 'likely<(↑ XCOMP)>(↑ SUBJ)'
(↑ SUBJ) = (↑ XCOMP SUBJ)

probable A

(↑ PRED) = 'probable<(↑ COMP)>(↑ SUBJ)'
(↑ SUBJ FORM) = il
(↑ COMP COMPL) = que
(↑ COMP TENSE) = future

(τ↑ PRED FN) = probable
(τ↑ COMP) = τ(↑ XCOMP)

Figure 21.6

The within-clause functions of the relativized phrases in the source and target language are determined by predicates which may be arbitrarily deeply embedded, but the relativized phrase in the target language must correspond to the one in the source language.

Let us assume that relative clauses can be analyzed by the following slightly simplified phrase structure rules, making use of functional uncertainty (see Kaplan and Maxwell 1988b for a technical discussion of functional uncertainty) to capture the non-local dependency of the relativized phrase (equations on the head NP are ignored):

(12) $\text{NP} \rightarrow \text{NP} \quad \text{S}'$
$(\uparrow \text{RELADJ}) = \downarrow$
$\text{S}' \rightarrow \quad \text{XP} \quad \text{S}$
$(\uparrow \text{REL-TOPIC}) = \downarrow \uparrow = \downarrow$
$(\uparrow \text{XCOMP}^* \text{ GF}) = \downarrow$

We can achieve the desired correspondence between the source and the target by augmenting the first rule with the following transfer equations:

(13) $\text{NP} \rightarrow \text{NP} \quad \text{S}'$
$(\uparrow \text{RELADJ}) = \downarrow$
$\tau(\uparrow \text{RELADJ}) = (\tau\uparrow \text{RELADJ})$
$\tau(\downarrow \text{REL-TOPIC}) = (\tau\downarrow \text{REL-TOPIC})$

The effect of this rule is that the τ value of the relativized phrase (REL-TOPIC) in the source language is identified with the relativized phrase in the target language. However, the source REL-TOPIC is also identified with a within-clause function, say OBJ, by the uncertainty equation in (12). Lexical transfer rules such as the one given in (8) independently establish the correspondence between source and target within-clause functions. Thus, the target within-clause function will be identified with the target relativized phrase. This necessary relation is accomplished by lexically and structurally based transfer rules that do not make reference to each other.

Differences in control. A slightly more complex but similar case arises when the infinitival complement of a raising verb is translated into a finite clause, as in the following:

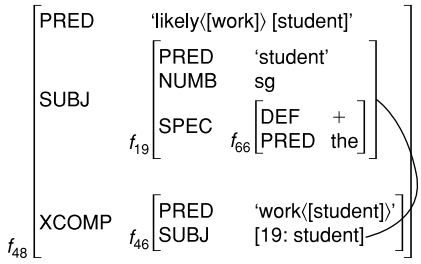


Figure 21.7

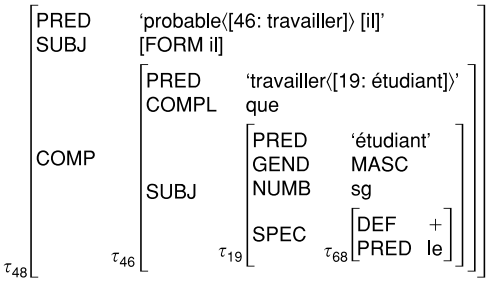


Figure 21.8

(14) (a) The student is likely to work.
(b) Il est probable que l'étudiant travaillera.

In this case the necessary information is distributed in the following way over the source, target, and transfer lexicons as shown in figure 21.6. Here the transfer projection builds up an underspecified target structure, to which the information given in the entry of *probable* is added in the process of generation. Ignoring the contribution of *is*, the f-structure for the English sentence identifies the non-thematic SUBJ of *likely* with the thematic SUBJ of *work* as follows (15, figure 21.7).

The corresponding French structure in (16) contains an expletive SUBJ, *il*, for *probable* and an overtly expressed SUBJ for *travailler*. The latter is

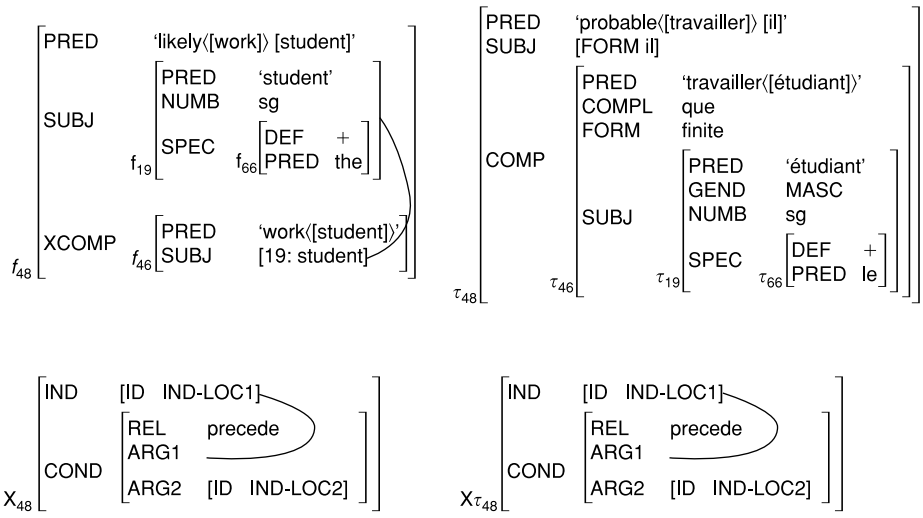


Figure 21.9

introduced by the transfer entry for *work* (16, figure 21.8).

Again this f-structure satisfies the transfer description and is also assigned by the French grammar to the target sentence.

The use of multiple projections. There is one detail about the example in (14) that needs further discussion. Simplifying matters somewhat, there is a requirement that the temporal reference point of the complement has to follow the temporal reference point of the clause containing *likely*, if the embedded verb is a process verb. Basically the same temporal relations have to hold in French with *probable*. The way this is realized will depend on what the tense of *probable* is, which in turn is determined by the discourse up to that point. A sentence similar to the one given in (13a) but appearing in a narrative in the past would translate as the following:

(17) Il était probable que l'étudiant travaillerait.

In the general case the choice of a French tense does not depend on the tense of the English sentence alone but is also determined by information that is not part of the f-structure itself. We postulate another projection, the temporal structure, reached from the f-structure through the correspondence χ (from $\chi_{\rho o \nu \iota \kappa \acute{o} \varsigma}$, temporal). It is not possible to discuss here the specific characteristics of such a structure. The only thing that we want to express is the constraint that the event in the embedded clause follows the event in the main clause. We assume that the temporal structure contains the following information for *likely-to-V*, as suggested by Fenstad et al. (1987):

(18) *likely* V
$(\chi\uparrow$ COND REL$)$ = precede
$(\chi\uparrow$ COND ARG1$) = (\chi\uparrow$ IND$)$
$(\chi\uparrow$ COND ARG2 ID$)$ = IND-LOC2

This is meant to indicate that the temporal reference point of the event denoted by the embedded verb extends after the temporal reference point of the main event. The time of the main event is in part determined by the tense of the verb *be*, which we ignore here. The only point we want to make is that aspects of these different projections can be specified in different parts of the grammar. We assume that French and English have the same temporal structure but that in the context of *likely* it is realized in a different way. This can be expressed by the following equation:

(19) $\chi\uparrow = \chi^{\tau}\uparrow$

Here the identity between χ and χ^{τ} provides an interlingua-like approach to this particular subpart of the relation between the two languages. This is diagrammed in figure 21.9. Allowing these different projections to simultaneously determine the surface structure seems at first blush to complicate the computational problem of generation, but a moment of reflection will show that is not necessarily so. Although we have split up the different equations among several projections for conceptual clarity, computationally we can consider them to define one big attribute value structure with χ and τ as special attributes, so the generation problem in this framework reduces to the problem of generating from attribute-value structures which are formally of the

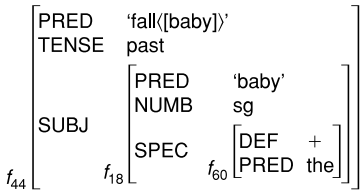


Figure 21.10

same type as f-structures (see Halvorsen and Kaplan (1988), Wedekind (1988), and Momma and Dörre (1987) for discussion).

Differences in embedding. The potential of the system can also be illustrated with a case in which we find one more level of embedding in one language than we find in the other. This is generally the case if a modifier-head relation in the source language is reversed in the target structure. One such example is the relation between the sentences in (20):

(20) (a) The baby just fell.
(b) Le bébé vient de tomber.

One way to encode this relation is given in the following lexical entry for *just* (remember that all the information about the structure of *venir* in French will come from the lexicon and grammar of French itself):

(21) *just* ADV $(\uparrow \text{PRED}) = \text{`just}\ \langle(\uparrow \text{ARG})\rangle\text{'}$
$(\tau\uparrow \text{PRED FN}) = \text{venir}$
$(\tau\uparrow \text{XCOMP}) = \tau(\uparrow \text{ARG})$

This assigns to just a semantic form that takes an AW function as its argument and maps it into the French *venir*. This lexical entry is combined with phrase-structure rule (22). This rule introduces sentence adverbs and makes the f-structure corresponding to the S node fill the ARG function in the f-structure corresponding to the ADV node.

(22) $\text{S} \rightarrow \text{NP} \qquad (\text{ADV})\ \text{VP}$
$(\uparrow \text{SUBJ}) = \downarrow\uparrow = (\downarrow \text{ARG})$

Note that the f-structure of the ADV is not assigned a function within the S-node's f-structure, which is shown in (23, figure 21.10). This is in keeping with the fact that the adverb has no functional interactions with the material in the main clause.

The relation between the adverb and the clause is instead represented only in the f-structure associated with the ADV node (24, figure 21.11).

In the original formulation of LFG, the f-structure of the highest node was singled out and assigned a special status. In our current theory we do not distinguish that structure from all the others in the range of ϕ: the grammatical analysis of a sentence includes the complete enumeration of ϕ-associations. The S-node's f-structure typically does contain the f-structures of all other nodes as subsidiary elements, but not in this adverbial case. The target structures corresponding to the various f-structures are also not required to be integrated. These target f-structures can then be set in correspondence with any nodes of the target c-structure, subject to the constraints imposed by the target grammar. In this case the fact that *venir* takes an XCOMP which corresponds to the ARG of *just* means that the target f-structure mapped from the ADV's f-structure will be associated with the highest node of the target c-structure. This is shown in (25, figure 21.12).

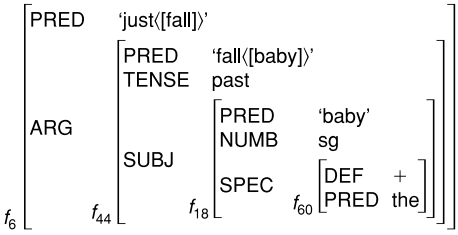


Figure 21.11

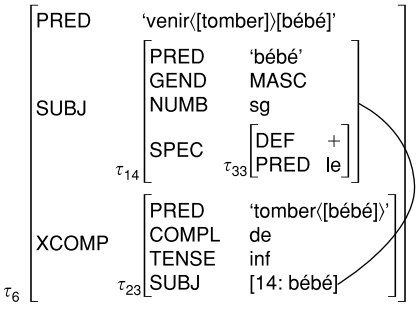


Figure 21.12

The above analysis does not require a single integrated source structure to map onto a single integrated target structure. An alternative analysis can handle differences of embedding with completely integrated structures. If we assign an explicit function to the adverbial in the source sentence, we can reverse the embedding in the target by replacing (22) with (26):

(26) $\text{S} \rightarrow \text{NP} \qquad (\text{ADV}) \qquad \text{VP}$
$(\uparrow \text{SADJ}) = \downarrow$
$\tau\uparrow = (\tau\downarrow \text{XCOMP})$

In this case the embedded f-structure of the source adverb will be mapped onto the f-structure that corresponds to the root node of the target c-structure, whereas the f-structure of the source S is mapped onto the embedded XCOMP in the target. The advantages and disadvantages of these different approaches will be investigated further in Netter and Wedekind (forthcoming).

Conclusion

We have sketched and illustrated an approach to machine translation that exploits the potential of simultaneous correspondences between different levels of linguistic representation. This is made possible by the equality and description based mechanisms of LFG. This approach relies mainly on codescription, and thus it is different from other LFG-based approaches that use a description-by-analysis mechanism to relate the f-structure of a source language to the f-structure of a target language (see for example Kudo and Nomura, 1986). Our proposal allows for partial specifications and multi-level transfer. In that sense it also differs from strategies pursued for example in the Eurotra project (Arnold and des Tombe, 1987), where transfer is based on one level of representation obtained by transforming the surface structure in successive steps.

We see as one of the main advantages of our approach that it allows us to express correspondences between separate pieces of linguistically motivated representations and in this way allows the translator to exploit the linguistic descriptions of source and target language in a more direct way than is usually proposed.

Acknowledgments

Thanks to P.-K. Halvorsen, U. Heid, H. Kamp, M. Kay, and C. Rohrer for discussion and comments.

References

Arnold, Douglas, and Louis des Tombe. 1987. Basic Theory and Methodology in Eurotra. In S. Nirenburg (ed.), *Machine Translation: Theoretical and Methodological Issues*. Cambridge: Cambridge University Press, 114–135.

Fenstad, Jens Erik, Per-Kristian Halvorsen, Tore Langholm, and Johan van Benthem. 1987. *Situations, Language and Logic*. Dordrecht: D. Reidel.

Halvorsen, Per-Kristian. 1983. Semantics for Lexical-Functional Grammars. *Linguistic Inquiry*, 14 (3), 567–613.

Halvorsen, Per-Kristian. 1988. Situation Semantics and Semantic Interpretation in Constraint-base Grammars. *Proceedings of the International Conference on Fifth Generation Computer Systems*, Tokyo, Japan, 471–478.

Halvorsen, Per-Kristian, and Ronald Kaplan. 1988. Projections and Semantic Description. *Proceedings of the International Conference on Fifth Generation Computer Systems*, Tokyo, Japan, 1116–1122.

Isabelle, Pierre, and Elliott Macklovitch. 1986. Transfer and MT Modularity. *Proceedings of COLING 1986*, Bonn, 115–117.

Kaplan, Ronald. 1987. Three Seductions of Computational Pyscholinguistics. In Peter Whitelock et al. (eds.), *Linguistic Theory and Computer Applications*. London: Academic Press, 149–188.

Kaplan, Ronald. 1988. Correspondences and Their Inverses. Paper presented at the Syntax and Semantics Workshop, April, Titisee, FRG.

Kaplan, Ronald, and Joan Bresnan. 1982. Lexical Functional Grammar: A Formal System for Grammatical Representation. In Joan Bresnan (ed.), *The Mental Representation of Grammatical Relations*. Cambridge, Mass.: MIT Press, 173–281.

Kaplan, Ronald, and John Maxwell. 1988a. Constituent Coordination in Lexical-Functional Grammar. *Proceedings of COLING 88*, Budapest, 303–305.

Kaplan, Ronald, and John Maxwell. 1988b. An Algorithm for Functional Uncertainty. *Proceedings of COLING 88*, Budapest, 297–302.

Kay, Martin. 1984. Functional Unification Grammar: A Formalism for Machine Translation. *Proceedings of COLING 1984*, Stanford University, 75–78.

Kudo, Ikuo, and Hirosato Nomura. 1986. Lexical-Functional Transfer: A Transfer Framework in a Machine Translation System Based on LFG. *Proceedings of COLING 1986*, Bonn, 112–114.

Momma, Stefan, and Jochen Dörre. 1987. Generation from f-Structures. In Ewan Klein and Johan van Benthem (eds.), *Categories, Polymorphism and Unification*, Edinburgh, Amsterdam, 147–168.

Netter, Klaus, and Jürgen Wedekind. 1986. An LFG-based Approach to Machine Translation. *Proceedings of IAI-MT 86*, Saarbrücken, 197–209.

Netter, Klaus, and Jürgen Wedekind. in press. Transfer by Projection. Stuttgart: IMS.

Reyle, Uwe. 1988. Compositional Semantics for LFG. In Uwe Reyle and Christian Rohrer (eds.), *Natural Language Parsing and Linguistic Theories*. Dordrecht: D. Reidel, 448–479.

Wedekind, Jürgen. 1988. Generation as Structure Driven Derivation. *Proceedings of COLING 88*, Budapest, 732–737.

This page intentionally left blank

22
Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation

Christian Boitet

Introduction: Why Is the Pivot Approach Not Universally Used?

The pivot approach seems best suited to the construction of multilingual M(A)T systems, for obvious reasons of minimality and economy. The idea is to translate the input text into a pivot language, and then from this pivot into the target language. In a multilingual setting with n languages, only n analyzers and n generators have to be constructed, comprising $2n$ grammars and $2n$ dictionaries (which give monolingual information and translations into or from the pivot lexicon).

In the transfer approach, there is the same number of analyzers and generators, but $n(n-1)$ transfers must be added. They transform source interface structures into target interface structures, using $n(n-1)$ transfer grammars and transfer dictionaries. If the interface structures contain a deep enough level of linguistic description, the transfer grammars are very small: the transfer dictionaries represent the bulk of the cost of the $n(n-1)$ transfers, they may be large, and they are more difficult to construct than monolingual dictionaries.

However, the pivot approach has been followed in very few systems until the eighties, when several new projects revived this design. Why was it almost abandoned for more than a decade, and why don't all modern MT systems rely on it? The answer cannot be simplistic, because there are several kinds of pivots, several kinds of interface structures, and several kinds of situations, which we will call "$1 \rightarrow m$" or "$m \rightarrow 1$", if translation occurs from one language into the m ($=n-1$) others, or into one language only, and "$n \leftrightarrow n$" if there are many language pairs (at least $2m$, with $m > 1$).

I. Pure Pivot Approaches

A pure pivot contains no information relative to the peculiarities of expression in the source and in the target language. This means that:

- there is an independent *pivot lexicon*, made of *pivot lexical symbols* (*terms* and "semantic" features);
- all grammatical information is replaced by *pivot grammatical symbols:* there is a universal notation for determination, quantification and its scope, actualization (time/modality/aspect), thematization (theme/pheme/rheme), abstract sex and quantity replace morphological gender and number, etc.;
- the *pivot structure* combines lexical symbols, annotated by grammatical symbols, by using *pivot relational symbols* like argument places or semantic relations; their level of interpretation is at least that of Tesnière's *actants*, or of Fillmore's *deep cases*, which are thought to remain almost invariant across families of languages, whereas syntagmatic categories and syntactic functions (used in c-structures and f-structures) do not.

I.1 Pure Pivot Lexicons Are Challenging ...

According to J. I. Tsujii (1987), there are three kinds of "pure pivots": *interpretation languages*, *standard languages*, *and conceptual decompositions*. This classification concerns the three aspects of lexical, grammatical and relational symbols, but we may concentrate on the pivot lexicon for the moment.

1.1 ... But Specific of a Domain (Interpretation Language) If the texts to be translated refer to a fixed and restricted domain, and are of a well defined type, it may be possible to define a completely artificial language to describe them. This is illustrated by the TITUS system (Ducrot 1982), which is still in use and evolving at the Institut Textile de France. The lexical symbols stand for concepts in the textile domain, and the input languages are controlled in such a way that there is a one-to-one correspondence between their vocabulary and the set of lexical symbols. In such situations, input of the texts is best done in an interactive way, and storing in the pivot form.

While this approach leads to excellent results in some situations, it is apparently not possible to use it without controlling the input language. Another

problem is that the lexical symbols and the corresponding natural vocabulary must be reconstructed for each new situation: the pivot lexicon is not universal, while the grammatical and relational symbols may be.

1.2 ... Or Specific of a Language Group (Standard Language) A standard language is an existing or artificial natural language, like English or Esperanto. Taking an existing natural language as pivot necessitates double translations for all pairs of languages which do not contain the pivot. In human translation, this leads notoriously to a decrease in quality, as ambiguities and misunderstandings (misanalyses and mistranslations, in our case) may increase. No experiments have been conducted yet in practice, with English or any other natural language.

If an *artificial* language is chosen, like Esperanto in the BSO project, all translations are double, and the difficulty is augmented by the lack of sufficient technical vocabulary. There is an accepted mean figure of 50,000 terms in any typical technical domain. But, then, there is a very interesting aspect to this choice: if the project succeeds on a large scale, the vocabulary of Esperanto will have been developed in such a way that the Esperantist dream may finally come true, as Esperanto will become a transnational language really able to support all kinds of international communication, without any political prejudice.

In order to reduce the number of added ambiguities, the BSO project seems to bracket the Esperanto text with "invisible" parentheses. This amounts to using some kind of surface structure. It is not clear to the author whether those parentheses are labeled or not, and, if yes, how. In any case, this addition may be viewed as a first step towards the idea of using structural descriptors to compose two transfer-based systems (see III.2 below).

If a natural language is chosen, so goes Tsujii's argument, the approach is limited to the language group or the language family of the standard language (Germanic or Indo-European for English). The "idiosyncratic gap" between Indo-European languages and Japanese has been pointed out more than once by Japanese colleagues. In the case of Esperanto, the basic vocabulary has been taken from several language families, but the problem still exists, because a choice has been made in each case, making it unavoidable that many distinctions and ways of expression are left out.

It should perhaps be added that very simple concepts may be expressed with different degrees of precision by languages of the same group. For instance, *mur* in French has two translations in Italian, *muro* (the wall seen from outside) and *parete* (the wall seen from inside). With the same distinction, *wall* may be translated as *Mauer* or *Wand* in German. This is true of a considerable number of names for concrete objects or notions (like color, kinship, ...).

1.3 ... And Always Very Difficult to Construct (Conceptual Decomposition/Enumeration) It is always difficult to construct a vocabulary in a coherent way, even for a natural language. Institutional bodies labor to create or normalize terminology. In any technical domain, however, perhaps less than 10% of the terms are normalized. This difficulty appears obviously when using interpretation languages or standard languages for MT, but the effort is immediately beneficial for areas other than MT.

J. I. Tsujii calls the third kind of pivot lexicon "conceptual decomposition." This technique has been popularized by R. Schank and his school since the early seventies. The idea is to define a small set of *conceptual primitives* (about 20 in the first versions) and to decompose all lexical items of a language in terms of them, obtaining *conceptual dependency* (*CD*) *structures*.

Of course, while predicative elements are relatively easy to decompose in this way, this is not true of the vast majority of the vocabulary of a natural language. For example, how does one distinguish all types of natural noises, rocks, plans, or animals, with so few primitives? The associated CD graphs are certain to be enormous.

Even if neuropsychology some day comes up with a proven set of, say, 200 or 2,000 basic primitives, the objection remains. The obvious solution used by natural languages and by some current Japanese MT projects (Fujitsu's ATLAS, NEC's PIVOT, ODA's CICC project on Asian languages) is to use *conceptual enumeration* on top of conceptual decomposition. In theory, this would amount to giving names to some CD graphs, and to use them in the construction of other, more complex, CD graphs. In practice, it seems that the aforementioned projects simply give a name to any new concept encountered, like "wall outside" and "wall inside," together with a definition written in natural language, very much like in usual dictionaries.

Then, the notion of concept may be equated with that of "meaning" in usual dictionaries. Complex terms such as "road haulage" are identified as concepts when this is clear in the considered language,

or when their translation into another language of the system is not considered language, or when their translation into another language of the system is not compositional (''camionnage'' or ''transport par route'' for this example, in French).

There are at least three main difficulties in the construction of such conceptual lexicons:

- First, there is the *sheer size of the set of concepts* to be defined for any reasonably general MT application. The Japanese CICC project is said to already use more than 250,000 concepts.
- Second, the *construction process is non-monotonic*. When a new concept is created from a term of some language, it is necessary to revise the dictionaries of all $n-1$ other languages. For example, ''wall'' is a unique concept when only English and French have been treated. When Italian or German comes in, it must be split.
- Third, *it is difficult to look for an existing concept* if its name is difficult to guess. For example, suppose one is adding a new complex term, like ''pros and cons,'' in one of the dictionaries, and that no translation into another language is available (in a usual dictionary). The only solution seems to try tentative definitions and to ask some support system to perform an associative search based on partial matches to check whether the pivot lexicon already contains the appropriate concept ar not.

It would be an oversimplification to think that this approach is a mere extension of the interpretation language approach, because one tries to take the union of many domains/situations: it is much more linguistic in spirit. A main difference is the possibility, and even the necessity, of ambiguity (see again the ''wall'' example). Also, there is no pretence to formalize all domains in which the MT system will work, as this would imply the explicit use of formal representations like the CD graphs, augmented by general and specific facts and inference rules, etc.

The ambition of the projects based on the conceptual decomposition/communication approach is enormous, but so are the human and financial resources allocated to them. Outside the field of MT, these projects may give two very important byproducts:

- the international normalization of a considerable amount of technical terms;
- a kind a multilingual encyclopedia.

I.2 Pure Pivot Structure Loses Information ...
It is extremely rare that two different terms or constructions of a language are completely synonymous. Using a pivot language makes it unavoidable that information useful for quality translation will be lost. But perhaps this is justified in view of the economic advantage in $n \leftrightarrow n$ situations.

2.1 ... At the Lexical Level Translating through an interpretation language certainly reduces distinctions between terms of the natural vocabulary, but, considering the situations in which this approach is used, this is of no importance. As a matter of fact, the overall process consists in creating an internal representation of the messages to be generated through the use of a ''quasi-natural'' (strictly controlled) input language, or even menus, and then in expressing them in many languages. This is a problem of generation rather than of translation in the usual sense.

In the case of a standard language, the problem is real. Of course, it is always possible to translate a simple term of the source language by a compound term (''wall seen from outside'') of the standard language, and then again by a simple term of the target language. But this must be done with care, as nothing prevents the input text from using the unmarked word for ''wall'' in an expression such as ''wall seen from outside'': this should be indicated in the pivot representation, or else the translation will be inexact. Also, the natural temptation for dictionary writers is to imitate usual bilingual dictionaries and to translate both ''muro'' and ''parete'' by ''wall,'' making it impossible to recover the distinction if going from Italian to German through English.

Perhaps the only way not to lose in lexical precision is to reach the ideal state where a complete conceptual dictionary will have been constructed for all the terms used in the class of texts to be translated. This certainly calls for active international cooperation.

2.2 ... At the Lower Interpretation Levels (Style) With the pivot methods, one obtains paraphrases rather than translations, because it is impossible to produce the desired parallelism in style, as all trace of the surface expression is erased. For example, it is not possible to force the system to translate the English passive by the French reflexive, in some predetermined contexts (*many equations are solved by iteration → beaucoup d'équations se résolvent par itération*) and by the French impersonal in other cases.

This makes it impossible to aim at a rough translation of professional quality. Perhaps it is the price to pay for the automation of translation in $n \leftrightarrow n$ con-

texts. But this limitation might be alleviated by the construction of (monolingual) stylistic editors, with which it would be a simple matter to change a whole text or selected portions of it, from imperative to subjunctive (*do this → you/one should do this*) or from some tense to another, etc.

2.3 ... At Non-Universal Grammatical Levels Another severe problem with the pivot approaches is the "all-or-nothing" problem: no translation is possible if analysis has not produced a correct result in terms of schematic relationships, a very difficult task. If the size of the unit of translation becomes larger than one sentence, e.g. one or several paragraphs, it is almost certain that the result of analysis will not be complete, and hence that the majority of units will have to be translated as fragments.

II. Transfer Approaches

The transfer approach is very frequently used, because of the difficulties mentioned above, and perhaps because $1 \rightarrow m$ or $m \rightarrow 1$ situations occur more frequently than $n \leftrightarrow n$ situations. This means that the source interface structure produced by the analyzer, usually a tree or a graph, contains lexical and grammatical information attached to the nodes and/or the arcs, and has to be submitted to a lexical and to a structural transfer, the latter incorporating some *contrastive knowledge* of the given language pair.

Structural transfer is simpler if the level of *interpretation* obtained is higher. These levels are, in ascending order, those of *syntactic classes* (noun, verb, adjective ...), *syntagmatic classes* (nominal phrase, relative clause ...), *syntactic functions* (subject, object, attribute, circumstantial ...), *logical relations* (predicate, first argument, second argument ...), and *semantic relations* (possession, quantification, accompaniment, instrument, location, cause, consequence, agent, patient, beneficiary ...), the last two being sometimes not distinguished.

II.1 The Hybrid Approaches May Be Worse, Because the Square Problem Remains ...

In the hybrid approaches, the lexicon is that of the source or target language, while the grammatical and relational symbols are universal. To go from a source interface structure to a corresponding target interface structure, a unique phase of lexical transfer is used. This means that, for each pair of languages, a big transfer dictionary has to be constructed: the square problem remains.

The term "hybrid pivot" was coined by Shaumyan in the sixties. Perhaps "hybrid transfer" would be a better term, because the main difference between the two approaches lies in the presence or absence of transfer dictionaries, and not of transfer grammars. In honor of Shaumjan, we will however continue to use his term.

1.1 ... If the Lexicons Are Only Monolingual (CETA) The hybrid pivot technique was first tried by the Grenoble group (CETA) between 1961 and 1970. It was then abandoned for the transfer approach. Until 1983, no $n \leftrightarrow n$ situation appeared, so that the square problem was not really a hindrance. The results obtained seemed also to demonstrate that the quality limit was really higher than with the previous method.

B. Vauquois also recommended this approach for the EUROTRA project, although the situation was clearly $n \leftrightarrow n$. There are three main reasons for that. First, the project was initially designed to be a development effort, starting from existing state-of-the-art techniques, and the construction of an adequate pivot language seemed too far-fetched. Second, it was clear that the pivot approach would necessitate a very strong discipline, and the centralized building of the linguistic components. Third, it was felt that the system should produce the best possible translations, in order to demonstrate the superiority of the second-generation (2G) architecture over the first generation's (1G), at a time when the EC was beginning to use SYSTRAN binary systems in Luxemburg.

1.2 ... And Even If Some Part Becomes Universal (EUROTRA) In 1983, when EUROTRA was launched and became more research-oriented, the transfer approach was kept, no doubt for the second reason: as linguistic development was to be scattered in nine, then 11 countries (for seven, then nine languages), the development of a common pivot lexicon was not envisaged.

Now, with 72 language pairs to consider, and some results to produce with 20,000 terms in each language by the end of phase 3, the square problem looks ominous. To alleviate it, S. Perschke has recently proposed to use a kind of conceptual lexicon for the technical terms. The idea is to associate a unique number (e.g., 19875545) to each such term, in the analysis and generation dictionaries. Then, the transfer dictionaries would not contain entries for these numbers, which would remain invariant through transfer.

Apart from the fact that using numbers is more difficult than to use mnemonic names, and that the square problem stays for the general vocabulary (up to 50,000 terms?), the problem of normalization and centralized control crops up again. The effort to assign those numbers in a reliable manner would be an enormous project in itself.

II.2 Transfer Architectures Using *m*-Structures ... In all approaches, analysis may be sequential or integrated. In the first case, the unit of translation is analyzed at each level of interpretation, the result being the representation of the unit at the last level for which analysis was successful. In this case, several structural transfers must be provided, one for each level of interpretation, or else a certain percentage of the input will not be translated, or translated by default as unrelated fragments (the "all-or-nothing" problem again). Of course, transfer at the syntagmatic level may be quite complicated, while it is quite simple at the last two levels, even if the two languages considered pertain to very different families. For reasons of modularity of development, this technique has been chosen by the EUROTRA project, as it had been 20 years earlier by CETA.

The alternative to sequential analysis is *integrated analysis*. It consists in letting the levels of interpretation interact during analysis, and in producing a *multilevel structural descriptor* as a result. Such a technique has been proposed by B. Vauquois in 1974, and has since been used in all MT systems developed with CETA's methodology. All the computed levels are represented on the same graph, a "decorated tree" which geometry is obtained by a simple transformation from a (not necessarily projective) dependency structure. With this scheme, some semantic information may be used to disambiguate at the syntactic level, as in the following sentences:

John drank a bottle of beer.

John broke a bottle of beer.

When disambiguation is impossible on the basis of linguistic criteria, the ambiguity can be coded in the structure, as for:

John lost a bottle of beer.

Also, it becomes possible to treat large units of translation, several paragraphs long, without encountering the "all-or-nothing" problem. If analysis at the highest levels of interpretation does not give satisfactory results on some part of the unit, this part and only it, is transferred on the basis of the lower levels, which act as *safety nets* (Vauquois & Boitet 1985).

2.1 ... Allow to Reach a Higher Quality One must admit that our linguistic knowledge is very incomplete. For many years, for example, some renowned laboratories have been looking for a universal notation to represent the tense/aspect/modality triad. This goal has not yet been attained. Moreover, even if such a description were found, it is not at all certain that computational linguists would be able to *compute* it from the input texts. One example of this situation is given by semantic relationships, which even human experts cannot assign in a reliable way on arguments (strong complements) of predicates (this has been experimentally proven several times, in particular in the EUROTRA ETL-4 reports).

Hence, the fact that there are some "traces" of the source language expression in the source interface structure may be used by the structural transfers translating from this language to compute a good rendering in the target languages. The grammar writer incorporates here his general contrastive knowledge of a given language pair, plus, if possible, some "translator's tricks," thus improving the naturalness and idiomaticity of the rough translation.

2.2 ... May Be Preferable in $1 \rightarrow m$ Situations Finally, it must be emphasized that $1 \rightarrow m$ situations seem to be the most frequent when high quality translation is desired. This is the case for the majority of the big firms, which produce their documentation in one language and translate it into many others. The domains and typologies are fixed, and ... going through a pivot would just add 1 lexical transfer to the m needed, and, of course, necessitate the construction of the pivot lexicon.

III. Both Approaches for the Future?

Because there are so many development efforts in many countries, with both approaches being used, it is not very risky to guess that they will both endure, with perhaps some solution.

III.1 Pivot

1.1 Domain-Specific Pivots: New Applications? With the enormous development of CAD/CAM and expert systems, it is very probable that many situations will appear, in which some information or documentation could be directly generated from the knowledge base of the system. As techniques for the

generation of natural language texts from a conceptual representation begin to be well known, the main problem will be to design efficient tools to construct a large variety of "quasi-natural" languages for the man-machine dialog.

1.2 Conceptual Decomposition/Enumeration: a Challenge The Japanese have embarked on a very ambitious multilingual and conceptual dictionary project, coordinated by the Electronic Dictionary Research Institute (EDR). Large scale work has begun on Japanese, English, Chinese, Korean, Thai and Malay.

This remarkable initiative is a challenge to other countries, in particular in the EC, to join in a common effort to develop an entirely new kind of *multilingual conceptual data base*. In the far future, we may think of analogous efforts to develop multilingual textual and grammatical resources, with many potential applications.

III.2 Transfer

2.1 Conversion from First to Second Generation We have mentioned some situations in which it may be advisable to develop new MT systems with the transfer approach. There are also situations in which one would like to improve existing MT systems (e.g., SYSTRAN) by converting them from first generation (1G) to second generation (2G), without losing the enormous amount of lexical and contrastive knowledge encoded in the bilingual dictionaries.

This effort could entail the development of neutral multilingual/multipurpose integrated dictionaries (Boitet & Nedobejkine 1986), which would be a first step toward the future integration in multilingual conceptual dictionaries, by the addition of references from terms to concepts.

2.2 Composition in $n \leftrightarrow n$ Situations: The Structured Standard Language Approach Finally, the idea of composing transfer-based systems might give a solution to the square problem, without requiring the construction of a pivot lexicon. Let us explain this in more detail. The input to a generator is a *target interface structure* which is not in general the same as the *source interface structure* produced by an analyzer. This is because the final form of the text is not yet fixed (paraphrases are possible), because polysemies not reduced by the transfer may appear as a special type of enumeration, and because the transfer may transmit to the generator some advice or orders (relative to the possible paraphrases), by encoding them in the structure.

Our idea is simply to physically divide the structural generation phase into two successive steps, the first choosing a paraphrase and producing a *source interface structure for the target language*, and the second the surface tree passed to the morphological generator. Then, this intermediate result of the generation can be fed to any transfer from the generated language, and the number of transfer dictionaries and grammars in a multilingual transfer-based system can be drastically reduced. This approach might be called the *structured standard language approach*.

For instance, consider the nine languages of the European Community. They may be divided in three groups: four Romance languages (French, Italian, Spanish, Portuguese), four Germanic languages (English, German, Danish, Dutch), and Greek. Instead of constructing 72 transfers, it might be enough, for the beginning, to construct only 14 transfers, six between the groups, for example French $\leftrightarrow$ English, Greek $\leftrightarrow$ Italian, German $\leftrightarrow$ Greek and Greek $\leftrightarrow$ English, and four in each group, for example Portuguese $\rightarrow$ Spanish $\rightarrow$ French $\rightarrow$ Italian $\rightarrow$ Portuguese and Danish $\rightarrow$ German $\rightarrow$ English $\rightarrow$ Dutch $\rightarrow$ Danish (or any programmatically better arrangement). To translate from Spanish to Dutch, one would then use the Spanish $\rightarrow$ French $\rightarrow$ English $\rightarrow$ Dutch route.

If one were to insist on never having more than double translations, it would be possible to make one of the most important languages as "center" (we consciously avoid the term "pivot," which is already overloaded), and to get a complete multilingual system by constructing just 16 transfers.

Conclusion: *m*-Structures with Esperanto or Pivot Lexicon?

Although perfectly pragmatic, this last solution might seem politically unacceptable. If so, why not take Esperanto as the central language? There would be obvious advantages, from the Esperantist and political points of view, while the differences with the BSO design would not be very important:

- interface structures would be *m*-structures, which would increase the upper limit in quality, and perhaps help to offset the loss due to systematic double translation;
- the representation of a text transported by the network would contain the Esperanto text, as well as its *m*-structure. In uncompressed form, the size of an

m-structure is slightly less than four times that of the corresponding text, in the author's experience.

Another suggestion for the future could be to use the *m*-structure approach with a pivot lexicon. With this true *hybrid pivot approach*, there would be no transfer dictionaries, but there might be up to $n(n-1)$ transfer grammars, to handle the contrastive phenomena. In case some transfer grammar were absent or incomplete, transfer would occur by default, on the basis of the universal grammatical and relational symbols produced by the analyzers.

Acknowledgments

I would like to thank the organizers of this conference on the new directions in Machine Translation for having given me the occasion to prepare this communication. Also, thanks to Elizabeth White, who helped remove many grammatical errors from the first draft of this paper.

References

Boitet, Christian. 1976. Un essai de réponse à quelques questions théoriques et pratiques liées à la traduction automatique. Définition d'un système prototype. Thèse d'Etat, Grenoble.

Boitet, Christian. 1988. Software and Lingware Engineering in Modern MAT Systems. In I. S. Bàtori et al. (eds.), *Handbook for Computational Linguistics*. Berlin: de Gruyter.

Boitet, Christian, and Nikolai Nedobejkine. 1981. Recent Developments in Russian-French Machine Translation at Grenoble. *Linguistics*, 19, 199–271.

Boitet, Christian, and Nikolai Nedobejkine. 1986. Toward Integrated Dictionaries for M(a)T: Motivations and Linguistic Organization. In *Proc. COLING-86*, Bonn, 423–428.

Ducrot, Jean-M. 1982. TITUS IV. In P. J. Taylor et al. (eds.), *Information Research in Europe*. ASLIB, London. In *Proc. of EURIM 5 Conf.*, Versailles.

Guilbaud, Jean-Philippe. 1984. Principles and Results of a German-French MT System. Lugano Tutorial on Machine Translation.

Guilbaud, Jean-Philippe. 1986. Variables et catégories grammaticales dans un modèle ARIANE. In *Proc. COLING-86*, Bonn, 405–407.

Nomura, Hirosato, Shozo Naito, Yasuhiro Katagiri, and Akira Shimazu. 1986. Translation by Understanding: A Machine Translation System LUTE. In *Proc. COLING-86*, Bonn, 621–626.

Slocum, Jonathan. 1984. METAL: The LRC Machine Translation System. In *Lugano tutorial on Machine Translation*.

Tomita, Masaru, Jaime G. Carbonell. 1986. Another Stride Towards Knowledge-based Machine Translation. In *Proc. COLING-86*, Bonn, 633–638.

Tsujii, Jun-Ichi. 1987. What Is Pivot? In *Proc. of Machine Translation Summit*. Hakone: JEIDA.

Vauquois, Bernard. 1975. La traduction automatique à Grenoble. Document de linguistique quantitative no. 29, Dunod, Paris.

Vauquois, Bernard. 1979. Aspects of Automatic Translation in 1979. IBM-Japan, scientific program.

Vauquois, Bernard. 1983. Automatic Translation. In *Proc. of the Summer School "The Computer and the Arabic Language,"* ch. 9, Rabat.

Vauquois, Bernard, and Christian Boitet. 1985. Automatic Translation at GETA (Grenoble University). In *Computational Linguistics*, 11, no. 1, 28–36.

Vauquois, Bernard, and Sylviane Chappuy. 1985. Static Grammars: A Formalism for the Description of Linguistic Models. In *Proc. of the Conf. on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Colgate Univ., Hamilton, N.Y., 298–322.

Veillon, Gérard. 1970. Modèles et algorithmes pour la traduction automatique. Thèse d'Etat, Grenoble.

Witkam, Toon. 1987. Interlingual MT—An Industrial Initiative. In *Proc. of the Machine Translation Summit*. Hakone: JEIDA, p. 135–140.

This page intentionally left blank

23
Treatment of Meaning in MT Systems

Sergei Nirenburg and Kenneth Goodman

1 Introduction

At the methodological level, polemics about the state of the MT art are most often couched in terms of the differences and tensions between the transfer and the interlingua approaches. Over the years, a great deal of folklore has accumulated about the pros and cons of each of these MT paradigms. We believe that the discourse on this topic is not as organized as it should be. A number of claims and opinions on the subject have been made publicly. Unfortunately, most of them appeared only in prefaces and introductions to books and articles on MT. After mentioning the predominant methodological issue on page 1 and briefly identifying their own positions, MT authors typically plunge into descriptions of their own systems or models without further analysis of the methodological issues. Under these circumstances, it is not surprising that the methodological argument is not conducted at an adequate level of detail.

Even in discussions devoted specifically to the "transfer vs. interlingua" issue (such as, for instance, the panel on this topic at the UMIST 1989 MT workshop) many of the arguments remain too general and iconic. As a result, a discrepancy can be detected between the methodological beliefs held by MT practitioners and the actual (theoretical and practical) preferences and results in the field. Even at the methodological level, criticism is often detected at opinions that are not, in fact, held or defended by one's opponents.

Judgments about paradigms may differ depending on the specific profile of a system. MT is simultaneously an empirical discipline and a technological pursuit. Depending on the primary direction of research and development in a project, different criteria should be used to evaluate the utility and quality of systems developed in it. There are (1) production systems, (2) production system prototypes, (3) proof-of-concept systems which demonstrate the utility of a theoretical or descriptive approach to MT or a component process in MT (e.g., syntactic analysis, treatment of referential meaning, etc.) and (4) technological testbeds for producing MT systems (including specialized knowledge acquisition interfaces, debugging tools, control environments, etc.).

One must also distinguish between evaluations of particular projects and evaluations of entire approaches. If it is claimed that Project A used Approach X and failed, it does not necessarily follow that Approach X is bad. Similarly, the claim that Project B used Approach Y and succeeded does not in itself mean that Approach Y is superior. One reason for this caution is that large MT projects tend to feature elements from several MT paradigms. Therefore, it is often a gross generalization to call a particular project purely interlingual or purely transfer. A finer-grain taxonomy of MT approaches is needed. This is a central methodological point of this paper.

Actually, the components of an interlingua text are produced by and informed by several interconnecting subsystems. In our knowledge-based MT system, KBMT-89 (Goodman and Nirenburg 1989), the interlingua is created by an analyzer that consists of a set of programs and knowledge sources, including source-language lexicons and grammars, mapping rules for syntactic features and structures, and an ontology or domain model. The generation side of the triangle is equally complex.

We would like to argue that the ultimate point of contention in methodological debates among the MT researchers is not so much the differences between the transfer and interlingua approaches but rather the attitude to treatment of meaning. It so happens that, as a rule, those MT workers who de-emphasize the importance of meaning extraction tend to favor transfer-oriented systems, while those who insist on understanding as a prerequisite of translation tend to prefer interlingua-oriented ones. The reasoning leading to this latter preference can be clarified using an example of a research program in meaning-oriented MT, namely KBMT-89.

In very general terms, our research-and-development activity can be characterized as follows. Its

methodological basis is meaning-oriented MT in an interlingua paradigm.[1] The Center for Machine Translation's KBMT systems are research-oriented systems that come under the rubrics "proof of concept" and "technological testbed." A great deal of attention is paid to the functionality range of software engineering, including architecture and control, as well as to the massive task of knowledge acquisition. A new version of the system is developed, tested and demonstrated, on average, annually. The systems we build and demonstrate are gradual approximations of an ideal interlingua MT system. At present, although some facets of our systems are relatively complete and stable (some of the grammars, parsers, the integrating control structure), we have only partially accounted for many others, such as, for instance, many areas of domain knowledge, the lexis of the languages involved and some of the heuristic rules ("microtheories") used for treating particular linguistic phenomena. Systems developed at CMT differ mainly in the amount of knowledge that has been accumulated for use in them. They share a number of important characteristics.

The flow of control in these systems is as follows. The input text is processed by a battery of text analysis programs. Using the knowledge recorded in the SL grammar and lexicon, these programs (after several stages of processing) produce an expression (usually called *interlingua text* or ILT) in a specially defined textual meaning representation language. Elements of ILT are produced based on lexical, grammatical and pragmatic meanings extracted from the source text. Some ILT elements are instances of concepts in a domain model. Some others are values of various semantic and pragmatic properties suggested as necessary components of ILT. Very generally,[2] ILTs are hierarchical structures of clause-level representation units connected through domain and textual relations from a predefined set and characterized by speaker attitude values.

At present, we cannot produce a complete ILT fully automatically. Therefore, our systems use an interactive "augmentor" (the concept of such a program was first demonstrated by Kay 1973). As our knowledge about language processing grows, we expect the role of the augmentor to diminish.

A complete ILT produced jointly by the analyzer and the augmentor is passed on to the generator suite of programs, which includes a text planner, a lexical selection module and a syntactic realizer. The generator uses a TL lexicon and grammar and other heuristic knowledge sources necessary for generation.

The number and nature of microtheories used for extracting and representing meaning components varies among system instances in our project. Microtheories are not necessarily "homegrown." They can be routinely imported and adapted. Thus, some of the heuristics can, in principle, depend on some surface phenomena in the source text. If such heuristics are used, they constitute a "transfer" element in a generally interlingual system.[3] We have no objection in principle to a transfer-oriented system as long as it incorporates a meaning treatment module. It seems, though, that at least for purely technological reasons it is simpler to formulate polysemy resolution rules in terms of domain model elements rather than lexical units of the source and target language, as is the practice in transfer systems that deal with polysemy resolution at all (e.g., SPANAM, Vasconcellos and Leon (1985)).

Meaning-based MT offers additional scientific incentives. It is a paradigmatic task for computational linguistics in that components necessary for a knowledge-based MT system are also necessary components in practically every other application system dealing with automatic processing of natural language. MT is, therefore, one of the most attractive comprehensive applications for various computational-linguistic theories—morphological, syntactic, semantic and pragmatic. Historically, "pure" transfer systems have not addressed the problem of semantic and pragmatic ambiguity resolution as a central problem for machine translation. These approaches are based on recognizing meaning without representing it (other than in terms of target language lexical units). Methodologically, this is a result of (1) expecting very small amounts of ambiguity and (2) relying on similarities between the source and target languages to try to "preserve" in translation any ambiguities that appear in the source text. When (or, in fact, if) a transfer system has any theoretical connections, they are predominantly connections to theories of syntax. Of course, such theories constitute the bulk of theoretical work in modern computational linguistics. But this state of affairs does not make analysis of meaning less essential. On the contrary, it becomes even more important for semantic theorists to have an adequate testbed for their ideas. And meaning-oriented MT provides just that.

The purpose of the above discussion of our view of meaning-based MT is not to give a complete or adequate account of its workings. As such, it certainly fails. Our intent is to illustrate the complexity of system organization, which cannot be readily sum-

marized in the triangular icon of the interlingua approach. In a narrow sense, the "interlingua" in our systems is the set of all well-formed expressions (ILTs) in the textual meaning representation language. In a broader sense, it should also include the domain model underlying (most entries in) the lexicons. It is in terms of the expressive power of this interlingua (and the feasibility of the "microtheories") that the quality of our version of the meaning-based approach to MT should be discussed. Other views of the treatment of meaning in MT should be judged on the quality of their own representations and analyses.

In what follows, we discuss several opinions that are frequently put forward in arguments about meaning oriented MT. The following list summarizes these opinions and puts them in a logical chain of arguments which goes from extremely strong and general criticisms toward more specific and limited ones. The list is by no means complete. We hope only that it is representative. After presenting the list of criticisms, we will evaluate each in turn and in greater or lesser detail.[4]

1. Translation is *not possible*; if it is, then

2. Meaning is *not required* for translation; if it is, then

3. Meaning is *not definable*; if it is, then

4. Meaning in different languages is different and *not compatible*; if it is compatible, then

5. One *cannot represent* this meaning in a language-independent way:

• the language of representation will be heavily slanted toward one particular natural language;

• it is difficult to come up with the necessary set of language-independent primitives and to ensure completeness of meaning representation. Furthermore,

6. It is not possible to base meaning representations on a complete logical calculus. Therefore, one can *never prove* the correctness of any representation, particularly that it is free of contradiction; or that the same meanings will be always represented similarly. If constraints of this sort are demonstrated to be manageable or unnecessary, then

7. It is impossible to ensure that the meaning can *actually be extracted* from the source-language text and rendered in the representation language; at least, it is not possible to extract meaning *completely automatically*.

8. Even if meaning-based translation systems can be built, they will produce not translations but rather paraphrases of source language texts.

2 Possibility of Translation

A long and rich history attaches to philosophical arguments against the possibility of translation as a re-representation of meaning. Of course the *locus classicus* is Quine's *Word and Object* and his theory of radical translation. The point of radical translation was that

> Manuals for translating one language into another can be set up in divergent ways all compatible with the totality of speech dispositions, yet incompatible with one another. In countless places they will diverge in giving as their respective translations of a sentence of the one language sentences of the other language which stand to each other in no plausible sort of equivalence however loose. (Quine 1960:27)

Quine's behavioristic "translation manuals" may be understood loosely as analogues of the grammars, lexicons and programs of machine translation. The idea of radical translation was not that it is impossible to translate natural languages—humans do it all the time—but that what is translated is not the "same meaning." In other words there are no meanings *qua* meanings to translate: The alleged absence of independent identity conditions for meanings entails that there are no language-neutral semantic entities. We examine this and related philosophical issues in Goodman and Nirenburg (1990), and will not develop the points here.[5] Suffice it to say that following Katz (1988) we believe (1) Quine was just mistaken and (2) it is not in any event clear how to apply his arguments to machine translation. The goal of Quine and some of his allies is mainly to demonstrate the underdetermination of scientific theories by evidence and it would be specious to transport the issues and arguments too quickly to our domain.

To contend that translation is not possible, then, is on one reading just false—and should be uncontraversially so. On another reading it entails, in some cases informally, a group of critiques of the "meaning" relation; and so it is to those critiques that we turn next.

3 Understanding and Translation

It has been suggested that meaning is not required for machine translation. The idea is that a source-language sentence might be translated automatically into a target-language sentence by *statistical* means. The idea is as old as MT itself and attracted Warren Weaver in the 1940s and informed the early approaches at RAND and the National Bureau of

Standards through the early 1960s (see de Roeck 1987 and references cited therein).

Most recently Brown et al. (1988) report on experiments with a statistical approach to machine translation which "... eschews the use of an intermediate mechanism (language) that would encode the 'meaning' of the source text." The contention in this approach is that "... translation ought to be based on a complex glossary of correspondence of fitted locutions" and more fully,

> Translation can be somewhat naively regarded as a three stage process:
>
> 1. Partition the source text into a set of fixed locutions.
> 2. Use the glossary plus contextual information to select the corresponding set of fixed locutions in the target language.
> 3. Arrange the words of the target fixed locutions into a sequence that forms the target sentence.

In other words, language in this approach is treated not as a productive system but as a fixed and unproductive set of canned locutions.

The applicability of an MT system built according to this approach is restricted to the cases where there are vast textual corpora of translation equivalents. But even when such materials are available, completely uninterrupted comparison will lead to errors simply because the human translators who produced the translations in the corpus in the first place do not translate word-for-word or even sentence-for-sentence. The meaning expressed by a lexical unit in the source language can be rendered as an affix or as a syntactic construction in the target language. Nagao (1989:6–7) writes:

> ... although they are infrequently used in European languages, in Japanese there are many words of respect and politeness which reflect the social positions of the speakers, as well as distinctly male or female expressions which lie at the heart of Japanese culture. These are factors which must be considered when translating between Japanese and European languages.... Even if those factors are not explicitly expressed in the target language, they should be inferable from the context, from the psychological state of the speaker, or from the cultural background of the language.

It will be difficult for a purely statistical system to detect such phenomena.

A major shortcoming of the statistical approach is as follows. What, that is, does one do when in a certain text the English word *lead* is translated into Russian as *provod* ("cable") 17 times and as *svinets* (the metal) 6 times? Can we, indeed, be democratic and go with the greater number of votes? Clearly not. Therefore, according to the statistical approach, one has to make step 2 in the above definition of the translation process a conditional one. The conditions will have to be formulated in terms of the "contextual information," that is, in terms of lexical units as such, syntactic structures, or lexical and other meanings. Depending on the particular choice from this list, the statistical approach will, we suggest, rediscover direct, transfer or interlingua models of MT.

While it does not seem that a purely statistical approach is adequate to the task of MT, we believe that a statistical component can be very useful in a practical MT system, both as an aid in knowledge acquisition and as a way of testing meaning preferences.

That meaning understanding is not necessary is also maintained by another group of researchers who observe that, for instance, the polysemous Spanish noun *centro* is translated into German as *zentrum* no matter which of the senses of *centro* was used in the SL text (see below). The question then is, why waste time detecting and representing the meaning of the input string when the target language correlate is always the same? Similar claims have been made about syntactic ambiguities (e.g., Pericliev (1984)) and ambiguities of prepositional phrase attachment (e.g., Kay (1989)).

A typical formulation of this position is given by Ben Ari et al. (1988:2): "It must be kept in mind that the translation process does not necessarily require full understanding of the text. Many ambiguities may be preserved during translation (Pericliev 1984), and thus should not be presented to the user (human translator) for resolution."

Similarly, Isabelle and Bourbeau (1985:21) contend that

> Sometimes, it is possible to ignore certain ambiguities, in the hope that the same ambiguities will carry over in translation. This is particularly true in systems like TAUM–AVIATION that deal with only one pair of closely related languages. The difficult problem of prepositional phrase attachment, for example, is frequently bypassed in this way. Generally speaking, however, analysis is aimed at producing an unambiguous intermediate representation.

This position is, in fact, a system-completeness argument. What it says is that, for a given SL-TL pair and (i) a given set of dictionary senses of each SL word and (ii) recognized SL syntactic patterns, there will be cases in which all the senses of a SL lexical unit will be realized by a single lexical unit in the TL, or an SL syntactic construction can be re-created without change in TL. The familiar sentence

(1) I saw a man on the hill with a telescope.

can be translated into some languages without the need to understand the dependency characteristics of the prepositional phrases, just by stringing them, in their original order, after the direct object. This type of knowledge allows the system builders to keep the sizes of lexicons and grammars smaller.

The set of arguments about preserving ambiguity has serious limitations. It is clear, for instance, that *centro cuidad* should be translated in colloquial English not as "town center" but as "downtown." Also, it is only possible correctly to render (1) in, say, Russian, if one understood the prepositional attachments. Otherwise it will not be possible to select prepositions and casual forms adequately. The argument in Pericliev (1984) is supported by a manual analysis of 200 short English phrases and their translations into Bulgarian. No syntactic ambiguity was found in about 150 of these phrases. In about 25 cases ambiguity could be preserved by simple substitution. And the remaining 25 could not be treated this way. Therefore, Pericliev's claims lack generality.

Considering the amount of work required to put together a non-trivial MT system, it is quite reasonable to strive to constrain the size of the knowledge acquisition task. At the same time, one must remember that a system strongly relying on the "ambiguity preservation" method is extremely vulnerable in situations where (1) the lexicon is growing while the system is in use or (2) when additional languages must be introduced. Every new word sense added to the lexicon carries the potential of ruining the possibility of retaining ambiguity in translation. And this means that extra attention must be paid to the maintenance of the lexicons.

The problem of working with increasingly large dictionaries and grammars remains to be solved for all MT systems, irrespective of the theoretical approaches they follow. There are also cases (especially when very concrete technological terminology is conceded) when knowledge about the field of translation and the authoring style of a particular type of text will lead to the possibility of rendering certain elements of SL through unconditional (and, possibly, multilingual) substitution by TL counterparts. However, actual MT systems will be judged by their "maximum" capabilities in treating complex, not simple, problems.

Some of the MT literature is devoted mostly to design and metalevel (not to say MT-theoretical) issues (King 1981, Arnold and des Tombe 1987, Warwick 1987). Typically, such contributions suggest or discuss an abstract theory of translation as a series of transformations among representations. A typical series of definitions follows:

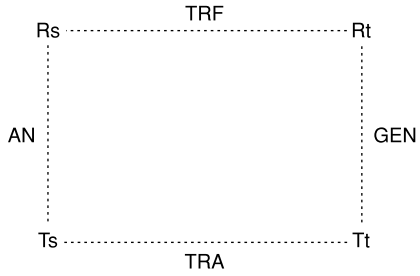


Figure 23.1
Here, Ts and Tt are texts, where a language is regarded as a set of texts. TRA is a binary relation, consisting of pairs of texts [Ts, Tt] where Tt is a translation of Ts. So, given two languages, SL and TL, TRA $\leq$ SL $\times$ TL. We introduce, furthermore, ρ which is a set of representations of some kind. Rs and Rt are both members of this set. We will write R when it is unimportant whether we are dealing with Rs or Rt, or when the context makes it clear which is intended (Johnson *et al.* (1985)).

Figure 23.1 forms a good basis for the study of representations in a transfer-based translation system:

In other abstract definitions of the translation process the number of transformations is larger, but in none of them is the question of semantic ambiguity dealt with centrally. Thus, for instance, having sketched an abstract view of the translation relation, Whitelock (1989:6) characteristically adds:

> One question I have not touched on here is the question of ambiguity. What I have been talking about is a many-many relation "possible translation" which may be computed monotonically from the axioms of the grammar and lexicon of the languages concerned. Optimally, this is viewed as a totally different question from determining the best translation, given an unbounded amount of real world knowledge, discourse context, etc. Computing this relation requires inference which is presumably defeasible.

Intrinsic in this statement is the opinion that what the understanding of meaning adds to the quality of translation is the possibility of getting to the "best" translation, as opposed, presumably, to "adequate" though not "best" one. We believe that the translation relation suggested by Whitelock has no way of guaranteeing even an "adequate" let alone the "best" translation. It has been amply and repeatedly demonstrated through multiple examples in the MT and natural language processing literature, starting at least with Bar Hillel (1960), that no adequate translation of realistic-size texts can be obtained if seman-

tic issues are not addressed. A great deal of ingenuity and *ad hoc* knowledge acquisition is needed to avoid ambiguity resolution in MT. And, in fact, translation using this approach can be successfully achieved only for carefully selected subsets of texts, served by dictionaries in which single-sense entries predominate. This is one reason why postediting is such a necessary stage in typical direct-approach and transfer-oriented MT environments. It seems that even when discussing general design issues in MT, it does not make sense to exclude considerations of ambiguity resolution.

4 Meaning across Languages

Some MT researchers adopt the position that different languages employ different concepts, or employ concepts differently, and this short-circuits attempts at meaning extraction. Thus Amano (1989:2) writes that "Natural languages have their own articulation of concepts according to their culture." To illustrate this point, Amano reports that where the English word *moustache* is customarily defined in English dictionaries as comprising hair on the upper lip, the Japanese *kuchi-hige* is defined in one (unspecified) Japanese dictionary as a "beard under the nose." (Actually, the ideographs for *kuchi-hige* stand for "lip" or "mouth" and "whiskers".) From this we are urged to infer that what Japanese speakers mean by *kuchi-hige* is somehow different than what English speakers mean by *moustache*. Of course, this opinion is simply a particularly hirsute version of Sapir–Whorfism that depends crucially on the vagaries of dictionary entries. Amano states that "natural languages have their own articulation of concepts according to their culture. Interlingua must naturally take account of this" (ibid.). But this is a misunderstanding of the concept of interlingua. What differs among languages is not the meaning representation but rather the means of realizing this meaning. The meaning of *kuchi-hige* and *moustache* will be represented in the same way in an interlingua text. The realizations of this meaning in the two languages will be different. It is in the interlingua-TL dictionary that a connection is established between an interlingual meaning representation and the language-particular linguistic expression.

This is not the place to argue against linguistic and cognitive relativism. The idea of linguistic relativity is, in fact, neutral with respect to the tasks of computational linguistics. It should be sufficient to point out that however convenient dictionaries might be as explicators of meaning for humans, it is a mistake to appeal to them as formal indexes of a culture's conceptual structure. That is to say, even within a language many terms may be rendered in different and apparently incompatible ways. To contend that meaning exists intralingually but not interlingually is to fall prey to such examples and to slip into the meanest sort of relativism, even unto idiolects. In practice, of course, indigenous *realia* can be described encyclopedically and then assigned a linguistic sign (possibly, a direct calque from the original language).

5 Feasibility of General Meaning Representation

One argument against language-independent meaning representation, usually referred to as interlingua, is known as "cultural-imperialist." To wit, the way the interlingua is built reflects the world view behind one dominant language. Examples of phenomena with respect to which "cultural imperialism" can be established include the cross-linguistic difference in subcategorization behavior of verbs, the grain size of concept description and the difference in attitude similar to the above *moustache* case. For instance, a single interlingua concept can be suggested to represent the main sense of the English *put* (as in *Put a book/glass on the table*). This might be considered a case of English cultural imperialism because in Russian this meaning can be expressed either as *polizit'* or *postavit'* depending on some properties of the object of *put*.[6] Additional examples abound in the MT literature.

The granularity of a large-scale meaning representation is always influenced by linguistic data, since the acquisition of knowledge necessary to support such a representation is done by humans who are, naturally, influenced by the languages they speak and the textual corpora and human-oriented dictionaries they use to determine meaning unit boundaries. It seems that the "cultural imperialism" argument is directed at the wrong target.

The simple view of the interlingua as a representation capturing all meanings in all languages is certainly limited because it talks about an ideal approachable only asymptotically. Compare, for instance, the following statement by Nagao (1989:6): "... when the pivot language method is used, the results of the analytic stage must be in a form which can be utilized by all of the different languages into which translation is to take place.... This level of subtlety is a practical impossibility." On a more technological level, Schneider (1989:128) justifies the choice of paradigm in the METAL project as follows:

METAL employs a modified transfer approach rather than an interlingua. If a meta-language were to be used for translation purposes it would need to incorporate all possible features of many languages. That would not only be an endless task but probably a fruitless one as well. Such a system would soon become unmanageable and perhaps collapse under its own weight.

This "maximalist" view of interlingua is so popular probably because it is conceptually the simplest. In operational terms, however, it is as useful to talk about such a conception of the interlingua as about a set of bilingual dictionaries among all the language pairs in the world. A practical interlingua should be viewed both as an object and as a process. Viewed as an object, developed in a concrete project, an interlingua should be judged by the quality of the translations that it supports between all the languages for which the corresponding SL-interlingua and interlingua-TL dictionaries have been built. As a process, its success should be judged in terms of the ease with which new concepts can be added to it and existing concepts modified in view of new textual evidence (either from new languages or from those already treated in the system). In practice, all interlingual systems start with the description of the semantic (sub)realms of a small set of languages and expand only when it becomes feasible from the standpoint of project resources. This is true about such different interlingual systems as ATLAS-II (Uchida 1989), Rosetta (Landsbergen 1989) or KBMT-89 (Goodman and Nirenburg 1984).

It is characteristic, though, that even interlingua-oriented workers find it necessary to offer qualifying explanations of their paradigmatic choices. Thus, Landsbergen (1989:85) writes:

1. From the point of view of the system's architecture Rosetta is clearly an interlingual system. It consists of an analysis component that translates from the source language into an intermediate language, or which the expressions are semantic representations, and a generation component that translates from this intermediate language into the target language.

2. On the other hand, the intermediate language of Rosetta is not a universal interlingua, but is defined for a specific set of languages. So Rosetta is not interlingual in this strict sense.

3. In an ideal interlingual system the analysis and generation component for each language can be developed independent of the other languages. We will not discuss here to what extent this is desirable or possible, but it is clearly not the case in Rosetta.

A system (such as Rosetta) based on the principles of meaning analysis and absence of direct correspondences between the elements of SL and TL must have the right to be called "interlingual" unapologetically. It does not seem appropriate (in fact, it looks like a double standard) to require completeness as proof of feasibility for interlingua, while allowing adequate behavior in a limited domain for a limited set of language pairs (usually, a single language pair) to be the criterion of success of a transfer system.

Often, the argument about infeasibility of interlingual MT is presented as a "basic assumption" and not argued for, as done, for instance, by Arnold and des Tombe (1987:117):

... the translation relation is fundamentally and irreducibly a relation between *linguistic* objects. The representation languages must be linguistic in nature, and cannot therefore be completely neutral with respect to different natural languages, in the way that a genuine interlingua would be.

It is not clear what is meant by "linguistic." Is the formalism Arnold and des Tombe suggest (or formalisms based on similar principles, presented, e.g., in Johnson et al. (1985) or Arnold et al. (1987)) in any sense a more "linguistic" notation than an artificial language designed to capture textual meaning? If "linguistic" is equated with "stemming from a syntactic theory" then we strongly disagree, because, in our understanding, translation is based on mapping meanings, not syntactic structures.

The "maximalist" view of the interlingua sometimes constitutes the main reason for not selecting this approach for a particular project. This is sometimes the case even in the presence of task specifications (such as multilinguality) which suggest an interlingua approach. Thus, the reasons for not selecting this approach for the original EUROTRA project are given by King (1981) as follows:

EUROTRA tries, at its deepest level of representation, to characterize the semantic relations between constituents in the text via a set of relations based on an expanded form of case grammar.... However, since the set of relations are defined as those useful for translation and are only "universal" within the project, there is no attempt to reach a [sic] ideal, genuinely universal semantic representation.

In reality, the EUROTRA approach evolved in such a way that many of the elements usually associated with the interlingua approach (first and foremost, analysis of meaning) are present in it (cf. Durand et al., forthcoming). Therefore, the traditional arguments against interlingua and for transfer

approaches should perhaps be presented today as arguments against the use of meaning in translation. If meaning is not considered essential for translation, then a version of a transfer approach should be the choice, since interlingual approaches crucially depend on meaning representation. We still believe that it would be more convenient and general to couch the meaning analysis in language-independent form rather than analyze the meaning of a source language in terms of lexical units of a target language (and this is what bilingual transfer dictionaries, in fact, do). But this argument is of a secondary nature. The main point established by this convergence in approaches is that treatment of meaning is central to the task of MT. We are, in reality, disagreeing only on ways and means.

We want at this point to discuss several well-known opinions about and criticisms of interlingual MT. It seems that most of them, indeed, refer to treatment of meaning rather than to interlingua as such.

The most well-known and large-scale early experiment with interlingual representations ended in self-admitted failure:

> ... we have tried an approximation of the interlingua ("pivot") approach and found it wanting. In the ... CETA system, the pivot representation was of a hybrid sort, using as vocabulary the lexical units of a given natural language, and as relations the so-called "universals" corresponding to our current logical and semantic relations, plus abstract features such as semantic markers, abstract time and aspect and so on. (Vauquois and Boitet, 1985:35)

Design characteristics of the CETA interlingua were, in fact, drastically different from those usually associated with interlingual systems. Hutchins (1986:190f) summarizes the characteristics of the CETA pivot language as follows:

> The formalism was designed primarily as an interlingua for syntactic features, i.e., as the common 'deep syntactic' base of the languages in the system.... Its lexicon, however, did not represent a common base; instead the pivot language conjoined the lexical units of whichever two languages were being processed.... In other words, while the CETA pivot language was a true interlingua in syntax, it was a bilingual "transfer" mechanism in lexicon. Further, it was not intended that all sentences with the same meaning would be analyzed as ... one unique pivot language representation. Nevertheless, although there were thus as many "pivot languages" as there were S-TL pairs analyzed, all shared the same syntax and in this respect CETA considered their formalism as a first step in the direction of a "universal language."

As described above, the design of CETA is interlingual only in name. In fact, it is practically identical to that of a standard modern transfer-based MT system!

Still, the fact that this system recognized itself as interlingual and self-admittedly failed has been used to justify objections to interlingual MT, for instance, in the METAL project:

> It is frequently argued that translation should be a process of analyzing the source language into a "deep representation" of some sort, then directly synthesizing the target language.... We and others (King, 1981) contest this claim.... One objection is based on large-scale, long-term trials of the "deep representation" approach by the CETA group at Grenoble.... After an enormous investment in time and energy, including experiments with massive amounts (400,000 words) of text, it was decided that the development of a suitable pivot language (for use in Russian-French translation) was not yet possible. (Bennett and Slocum 1985:112)

Comparing this opinion to the above discussion of the CETA project, one has to conclude that the self-admitted failure of CETA should have raised doubts about the feasibility of the transfer approach rather than the interlingua one.

It is sometimes claimed that meaning representation that does not use elements of natural language is difficult to design: "It is very difficult to design [a meaning representation] in the first place, and ever more so if the vocabulary must also be independent of any particular natural language" (Vauquois and Boitet, ibid.).

In the years since CETA was designed, a large body of knowledge has been acquired in the area of representing models of real-world entities in the computer. And even though the task still remains difficult it is more feasible using the modern knowledge representation languages, advanced knowledge acquisition interfaces with built-in consistency and validity checks, suites of programs for processing machine-readable human-oriented dictionaries and encyclopedias, etc. With respect to the choice of names for primitives (the "vocabulary" of Vauqouis and Boitet), different knowledge-based systems choose different approaches (e.g., in KBMT-89 the primitives have the status of elements in an artificial language, while in the PREMO system (Slator and Wilks (1989)) English word senses are used).

Another typical criticism of meaning-based MT, expressed as a criticism of the interlingua approach, concerns the process of TL text generation. Vauquois

and Boitet (ibid.) write: "The absence of surface-level information makes it impossible to use contrastive knowledge of two languages to guide the choice between several possible paraphrases at generation time." This opinion is seconded by Warwick (1987:28):

> One major difficulty with the interlingual approach—aside from the complexity of defining such an abstract model—was that language-specific attributes necessary for defining translation equivalents on the lexical and structural level were neutralized in the interlingual representation, thereby complicating the task of generation considerably.

In a typical transfer system TL generation usually is concerned only with the syntactic part of the process. Text planning and lexical selection are both avoided, the former by uniformly translating every SL sentence by a sentence in the TL,[7] the latter by substituting TL lexical units through bilingual dictionaries. In fact, in early versions of transfer systems generation was little more than a left-to-right scanning and writing out of the terminal elements in a transfer phrase structure tree.

As long as lexical ambiguity is not treated in an MT system, the traditional absence of real lexical selection mechanism is justified simply because there isn't any choice—a single translation variant is suggested for every SL lexical unit. If a more sophisticated variety of the transfer approach can incorporate lexical ambiguity resolution while continuing to use TL as the language for representing the meaning of SL lexical units, then lexical selection in generation may continue to be a non-problem for approaches which use contrastive lexical knowledge. The crux of the matter is, however, still on the analysis side. By using a metalanguage with a higher expressive power than a natural language (we are talking about expressive power for computer programs, not humans!) a meaning-oriented MT system can allow lexical selection in generation to be performed at the level of sufficiently fine-grain semantic features, not monolithic lexical units.

This allows one to smooth out many cross-linguistic incompatibilities, such as problems of inexpressibility of certain concepts in single-word lexical units in some languages. Multiple examples of such phenomena can be found in MT and general linguistic literature, many of them dealing with translation of kinship terms.[8]

Yet another objection to the interlingua approach to MT is based on "practical" considerations. Bennett and Slocum (*op. cit.*) contend that

> since it is not likely that any NLP system will in the foreseeable future become capable of handling unrestricted input—even in the technical area(s) for which it might be designed—it is clear that a "fail-soft" technique is necessary. It is not obvious that such is possible in a system based solely on a pivot language.

"Fail-softness" is a worthy goal for a software system. However, this concept is invoked in the MT literature usually and only to stress a theoretical point, as in the passage quoted just above. In practice, neither transfer-based nor interlingua-based systems have at present a good means of dealing with unexpected or ill-formed input. Nothing in knowledge-based MT per se precludes the design and implementation of architectures and algorithms facilitating fail-softness. Just as in transfer systems a target lexical unit can be picked at random (or based on probabilistic judgments, which amounts to the same thing) when no disambiguation is possible, so in interlingua systems some decisions could be made based on similarly weak heuristics. The above criticism is, thus, a non-criticism. It probably stems from the observation that such weak heuristics are seldom used or discussed in meaning-oriented projects because these projects are typically research-oriented rather than devoted to building production system prototypes. However, if such a prototype is built using a meaning-oriented approach the objective of fail-softness can be achieved equally well.

6 How Formal Must Meaning Representation Be?

It is widely supposed that machine translation requires at ground a fully interpreted logical calculus, that a meaning-based approach cannot be presented with such formal rigor and hence that meaning-based MT cannot succeed. This argument may be understood as demanding formal proofs of the correctness of translated meaning representations. Without such proofs, it is supposed, there is no guarantee that a translation will be free of contradiction or that the same meanings will be always represented similarly.

The formalist approach to machine translation is heir to Montague's view that there is or should be no distinction in principle between natural and formal languages. But even if Montague, thus glossed, were correct, it would not follow that uniquely formal representations are necessary for the task of machine translation. That is to say, with Wilks (1989:3),

> ... we do need representations (as opposed to the current trend of connectionism ...), but their form, if interpretable, is largely arbitrary, and we may be confident it has little relation to logic. I shall restate the view that the key contribution of AI in unraveling how such complex tasks as "understanding" might be simulated by a machine lies not in representations at all but in particular kinds of procedures.... It would be the most extraordinary coincidence, cultural, evolutionary, and intellectual, if what was needed for the computational task should turn out to be formal logic, a structure derived for something else entirely. Although, it must be admitted, strange coincidences have been known in the history of science.

The demand for proofs that a target language text will contain no contradiction is of course a demand that cannot be met. But, fortunately, the problem of avoiding contradiction—in machine translation in particular and natural language processing in general—is an empirical issue and not clearly delimited by formalist claims and purported requirements. That is to say, while it might be nice to be able to offer such proof, it would be a grievous error to abandon any enterprise unable to provide a formal proof of its future success. Indeed, the formalist gambit has been tried against any number of sciences, including physics, and has come up short.

It is perhaps worthwhile to point to Quine's (1960) admission that the only things that can be radically translated are the logical connectives. It is not clear how one would press the point, but one might confront formalist demands by suggesting that if the connectives can be deterministically translated, then the (formal) avoidance of contradiction will not be quite so difficult as proposed. At any rate, the intuitions underlying "not," "and," "or" and so forth are indisputably common and accessible to natural-language users in the absence of any sort of formalism. If they can be formalized, so much the better for logic; but on what grounds is this formalization required for natural-language *understanding?*

The formalist claim is sometimes made by criticizing uninterpreted formalisms. The elements from which our representations are built are "interpreted" in terms of an empirically constructed domain model rather than through an axiomatically defined set of possible worlds or well-formed formulae in a logical system. To be sure, one must avoid over-facile appeals to future research and empirical criteria as a hedge against formalist strictures. Nonetheless, such a line can productively be deployed against the claim that meaning-based MT cannot ensure that same meanings will get the same translations. If sameness of intralingual meaning is in fact preserved in translation—as corroborated by the judgments of bi- or multi-lingual humans, say—then this should be regarded as evidence in favor of the meaning-based approach. It would be folly indeed to disregard such evidence in the absence of a formal proof of the possibility of such evidence!

7 Extractability of Meaning

It is argued that it is impossible to ensure that the meaning can actually be extracted from the source-language text and rendered in the representation language. As stated above, the present state of the art does not allow a completely automatic disambiguation and representation of all the semantic and pragmatic phenomena. This is especially true for systems like those coming out of the KBMT project at the Center for Machine Translation, in which the expected results of analysis are very detailed.

Hutchins summarizes the scene as follows (1987:49):

> In semantic analysis there has been successful treatment of homography and syntactic ambiguity; and there have been successful implementations of case frames, of semantic features, of distributional semantic information, and recently of Montague semantics; but, nevertheless, the profounder problems of interlingual semantic analysis have proved clusive.

These "profounder" problems presumably include treatment of reference (including ellipsis), abductive inference-making on the basis of word knowledge, speaker attitudes, indirect speech acts, stylistic factors, etc. We are making inroads into these and other different areas. In the meantime, the reliance on the concept of microtheories, the continued work on the acquisition of domain models and the use of an interactive argument (a program which supports the interactive editing functionality to treat those types of meaning which cannot be treated automatically within the current state of the art) make meaning-based systems feasible. This has already been demonstrated at the research level.

One of our tasks is to demonstrate the utility of this approach through a production system prototype. As we mentioned above, the role of the augmentor will progressively diminish as our research on meaning extraction progresses. But is strange to doubt that it is progressing.

8 Translation and Paraphrasing

Hutchins (1987:49) claims that in meaning-oriented MT systems "... the abstractness of 'content' representations results in losses of information about 'surface' structures of texts" and from this he concludes that "versions produced by AI methods are not translations but rather paraphrases."

This opinion relies too much on the formulation of translation as a relation among texts, not among textual meanings (cf. the similar definitions in Johnson et al. (1985) and Arnold and des Tombe (1987) as quoted above). If we agree that the invariant in translation is meaning, then translation becomes a paraphrase, only a special one; in this type of paraphrase the lexical, grammatical and prosodic means of a different language are used (see, e.g., Whitelock (1989) for a similar argument).

In fact, the "paraphrasing as translation" argument is a facet of a more general question: Does one need to treat the form of the input text during translation? This question naturally arises because the dichotomy of substance and form has been a central point of discussion in such fields as semiotics and art theory and history. By "form of text" we refer a number of diverse phenomena: the syntactic structure of the input sentences; its phonetic and prosodic properties, such as alliteration, meter, rhyme, etc.; and the layout of a printed page, which can include diagrams, formulas, pictures, examples and other highlighted material, special fonts and so forth.

The layout of the text on a page is a feature independent of text meaning but influences the overall impact of the text. It can be called on to carry an esthetic message (as, for instance, in Apollinaire's poems or Lewis Carroll's tale written in the *form* of a tail in *Alice in Wonderland*). In expository, esthetically neutral text, which is the type of text which is machine-translatable, it is sometimes desirable to preserve page layouts in translation (especially, for pages with diagrams, illustrations, etc.), as, for instance, in the case of multilingual equipment manuals.

It is clearly difficult to preserve phonetic characteristics of the source text in the target text, and not only for computers. We will therefore not expect to deal with these issues. However, the use of special fonts (e.g., italics) carries a meaning which has to be recreated in the target text. Sometimes lexical units from languages other than the main language of the text are highlighted. These should be recognized as material not to be translated but rather reproduced "as is" in the target text. However, in some cases italics are used for purposes of making sentential stress (e.g., "I do not want to see *any* of them"), and in such cases this meaning should be represented and later re-created in the target language using its own means of expressing sentential stress.

Outside the field of artistic texts—poetry and fiction—preservation of the syntactic form of the source text in translation is completely superfluous because the meaning and use of, say, passive voice constructions in a source and a target language should not necessarily be identical. Direct structural correspondences between certain pairs of languages can be exploited in MT systems of a particular type, but they should be treated as idiosyncratic occasions rather than phenomena that occur as a rule and should, therefore, be preserved in translation. From the point of view of quality of expository text translation, it is immaterial whether the syntax of the target sentences is similar to that of the source sentences.

To summarize, there is no reason to aspire to translate the form of the input text. However, if an MT system does not possess sufficient knowledge to analyze SL texts deeply enough to allow understanding sufficient for realization of corresponding TL texts, it may rely on preserving the syntax of the source text in the target text as a very crude decision-making heuristic, regularly violated in a large number of cases.

9 Conclusion

While we have so far emphasized the key points of difference between the two main MT paradigms, it will be productive to conclude with mention of the positions which seem to be held jointly by all MT system developers. These platforms for agreement seem to include the following:

- Translation is a relation between texts in the source and target languages, such that the invariant between them is meaning. In other words, translation is rendering a set of meanings realized in a source language using the realization means of a target language.
- MT deals with expository texts, where the artistic considerations do not play an important role.
- Meanings in such texts are, in practical terms, completely expressible in all relevant source and target languages.
- Fully automated MT is not feasible at present, but

• The main research direction is toward full automation.

Additionally, here are some positions that are held by researchers in meaning-oriented MT but are not emphasized by other MT workers:

• SL ambiguity resolution is the main technical goal to be achieved by MT systems.

• Paradigmatic and other design considerations must crucially take into account the above requirement.

Interlingual MT systems tend to favor the meaning-based approach, while transfer systems tend to render meaning without the added requirement of representing it. Theoretically, meaning-oriented MT is not restricted to the interlingua paradigm. One can in principle incorporate meaning analysis into the transfer approach. However, in practice, as such attempts proliferate, it will become clear that the interlingua paradigm is more convenient for the support of the analysis of meaning. We also believe that the amount and complexity of knowledge acquisition for interlingual MT systems is at worst roughly equal to that which would have to be mastered for meaning-oriented transfer MT. At best, the acquisition component of an interlingua approach will be more compact and well-organized.

This paper has dealt exclusively with conceptual arguments. There are a number of practical—technological and methodological—issues relating to the differences among MT approaches. We plan to discuss them in a separate paper.

Notes

1. This statement will immediately instantiate the familiar triangular diagram in the mind of the reader. It is worth remarking that this quasi-standard illustration of the interlingua system design (and the equally ubiquitous trapezoid or rectangular diagram of the transfer system design) can be confusing or misleading. The use of circles, arrows and labels to represent the top-level structure of a system should properly be understood as the creation of a sort of visual slogan. There is much detail that does not easily lend itself to a graphical representation. In simple sketches of interlingua systems, for instance, the source language is seen at the left and giving rise via an arrow to an "interlingua" that in turn points with an arrow to the target language. In simple sketches of the transfer architecture (cf. figure 23.1) it is quite difficult to express the actual nature of the transfer structures.

2. A detailed description of our text meaning representation language, TAMERLAN, is given in Nirenburg and Defrise, forthcoming.

3. One such heuristic is the decision to retain in the target text the boundaries of sentences in the source text. It is well known to human translators that translations can be improved if one can combine some of the source sentences together or break some of them into several sentences. If knowledge is available for judging when such actions are appropriate, a meaning-based MT system can use it to determine sentence boundaries in the target text. But if this microtheory is not yet available, a good working heuristic is to copy the boundaries in the source. Other phenomena are treated in meaning-based systems in ways similar to their treatment in transfer systems. One such phenomenon is the handling of unambiguous terminological lexis (such as, for instance, chemical nomenclature).

4. Discussions of points 2 and 3 are combined in Section 3 below. The other criticisms are considered in separate sections.

5. Note that our meta-argumentative chain appears in an abbreviated form in the companion paper to plot out philosophical issues related to the differences between the transfer and interlingua MT approaches.

6. The difference can be glossed as that between *put flat* and *put upright*. A book can be "put" either way; a glass will be usually "put upright."

7. This is a very good approximation of a general translation rule, but still constrains the expressive power of a generator. The ability and license to break SL sentences into several TL sentences or combine several of them into one is, as above, a powerful weapon in the hands of a human translator.

8. One of the latest contributions is Amano (1989) in which it is suggested that Japanese has two lexical units corresponding to the English word *aunt*, one referring to an older sister of a parent and another referring to a younger sister. (In fact, the two words are phonetically the same, though different Kanji characters are used to represent them.) Note that Amano uses this example to support his opinion that direct correspondences between languages alleviate the problem of lexical gaps of this sort. Indeed, his criticism of the interlingua approach includes the statement that, in the cases like the above, use of a descriptive phrase like "father's younger sister" constitutes explanation rather than translation. Following this logic, real translation, then, will necessarily involve either a meaning loss or a potential error in translation. Indeed, for translation from Japanese into English, if the "explanation" mode is to be avoided, both the Japanese lexical items will have to be rendered as "aunt" in English. This is meaning loss. Establishing correct correspondence in the opposite direction will be utterly impossible without extra knowledge (the relative age of the person in question and one of her brothers or sisters)—either using bilingual correlations or using the interlingua method. The difference is that a typical transfer MT system does not have a mechanism to support such an inference even if this knowledge is in principle available, whereas interlingual systems are in principle designed with such problems in mind.

References

Amano, S. 1989. On Interlingua Approaches—From the Point of View of *traduttori traditori*. Presented at the MIT Workshop "Into the 90s." Manchester.

Arnold, D., and L. des Tombe. 1987. Basic Theory and Methodology in EUROTRA. In S. Nirenburg (ed.), *Machine Translation: Theoretical and Methodological Issues*. Cambridge: Cambridge University Press, 114–135.

Arnold, D., S. Krauwer, L. des Tombe, and L. Sadler. 1987. "Relaxed" Compositionality in MT. *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation*. Pittsburgh, June 1988.

Bar Hillel, Y. 1960. The Present Status of Automatic Translation of Languages. In F. L. Alt (ed.), *Advances in Computers* (Volume 1). New York: Academic Press, 91–163.

Ben Ari, D., M. Rimon, and D. Berry. 1988. Translational Ambiguity Rephrased. *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation*. Pittsburgh, June 1988.

Bennett, W. S., and J. Slocum. 1985. The LRC Machine Translation System. *Computational Linguistics*, 11:111–121.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. A Statistical Approach to French/English Translation. *Proceedings of the Second International Conference on Theoretical and Methodological Issues in Machine Translation*. Pittsburgh, June 1988.

Durand, J., P. Bennett, V. Allegranza, F. van Eynde, L. Humphreys, P. Schmidt, E. Steiner. 1991. The EUROTRA Linguistic Specifications: An Overview. *Machine Translation*, 6:103–147.

Goodman, K., and S. Nirenburg. 1989. KBMT-89. CMU CMT Technical Report.

Goodman, K., and S. Nirenburg. 1990. To Save the Semantic Phenomena: Machine Translation and Interlingua Texts. Presented at the Fifth International Conference on Computers and Philosophy. Stanford, August.

Hutchins, W. J. 1986. *Machine Translation: Past, Present, Future*. Chichester: Ellis Horwood.

Hutchins, W. J. 1987. Prospects in Machine Translation. *Proceedings of MT Summit*. Japan.

Isabelle, P., and L. Bourbeau. 1985. TAUM–AVIATION: Its Technical Features and Some Experimental Results. *Computational Linguistics*, 11:18–27.

Johnson, R., M. King, and L. des Tombe. 1985. EUROTRA: A Multilingual System under Development. *Computational Linguistics*, 11:155–169.

Katz, J. J. 1988. "The Refutation of Indeterminacy." *The Journal of Philosophy*, 85, 227–252.

Kay, M. 1973. *The MIND System*. New York: Algorithmic Press.

Kay, M. 1989. Talk at MT Summit II. Munich. August.

King, M. 1981. Design Characteristics of a Machine Translation System. *Proceedings of IJCAI–81*, 43–46.

Landsbergen, J. 1989. The Rosetta Project. *Proceedings of MT Summit II*, Munich, 82–87.

Nagao, M. 1989. *Machine Translation: How Far Can It Go?* Oxford: Oxford University Press.

Nirenburg, S., and V. Raskin. 1986. A Metric for Computational Analysis of Meaning: Toward an Applied Theory of Linguistic Semantics. *Proceedings of COLING–86*, Bonn, 338–340.

Pericliev, V. 1984. Handling Syntactic Ambiguity in Machine Translation. *Proceedings of COLING–84*, 521–524.

Quine, W. V. O. 1960. *Word and Object*. Cambridge, Mass.: The MIT Press.

de Roeck, A. 1987. Linguistic Theory and Early Machine Translation. In M. King, ed., *Machine Translation Today*. Edinburgh: Edinburgh University Press, 38–57.

Schneider, T. 1989. The METAL System. Status 1989. *Proceedings of MT Summit II*, 128–136.

Slator, B., and Y. Wilks. 1989. PREMO: Parsing by Conspicuous Lexical Consumption. *Proceedings of the International Workshop on Parsing Technologies*. Pittsburgh, August, 401–413.

Uchida, H. 1989. ATLAS. *Proceedings of MT Summit II*. Munich, 152–157.

Vasconcellos, M., and M. Leon. 1985. SPANAM and ENGSPAN: Machine Translation at the Pan American Health Organization. *Computational Linguistics*, 11:122–136.

Vauquois, B., and C. Boitet. 1985. Automated Translation at Grenoble University. *Computational Linguistics*, 11:28–36.

Warwick, S. 1987. An Overview of post-ALPAC Developments. In M. King (ed.), *Machine Translation Today*. Edinburgh University Press, 22–37.

Whitelock, P. 1989. Why Transfer and Interlingua Approaches to MT Are Both Wrong: A Position Paper. Presented at MT Workshop "Into the 90s." Manchester.

Wilks, Y. 1989. Form and Content in Semantics. Manuscript, Computing Research Laboratory, New Mexico State University, Las Cruces.

Wilks, Y. 1988. Philosophy of Language and Artificial Intelligence, Technical Memorandum MCCS 88–132, Computing Research Laboratory, New Mexico State University.

This page intentionally left blank

24
Where Am I Coming From: The Reversibility of Analysis and Generation in Natural Language Processing

Yorick Wilks

Introduction

The two general issues discussed in this paper are:

1. the symmetry (or otherwise) of analysis and generation in natural language processing, and
2. from what structure should generation come?

These related questions are contained within the well-known fable Wittgenstein tells in his *Remarks on the Foundations of Mathematics*, where he imagines entering a room in which a very old man is counting "... 5, 4, 3, 2, 1, 0 ... whew"! Asked why he is so exhausted, the old man responds that he has just finished counting down the natural numbers from infinity. The question implied is why this story seems more absurd to us than a symmetrical story of someone starting off from zero upwards, since the two tasks are essentially the same?

Although both tasks are impossible, we can in this case spot the difference fairly quickly whereas, in the case of analysis and generation, the situation is hardly the same. Both these tasks are possible (we are all existence proofs of that) and we think we have a much clearer idea of what the starting representation is than in the number case. Moreover, few researchers think those two tasks are as asymmetrically reversible as in the fable.

There has been a revival of interest in the reversibility issue in recent years (see, for example, Jacobs 1988; Appelt 1989, Russell et al. 1990) but the roots are much further back. The original sources of symmetry and asymmetry in computational linguistics (CL) seem to be (a) fallacious Chomskyan arguments and (b) the long-standing tendency in CL to trivialize generation in a way that was impossible for analysis. Winograd's SHRDLU (1972) is a paradigm of the latter, as of so many other trends, with his heavyweight, knowledge-based analysis, followed by a fairly trivial word-in-slot pattern-based generation. It is hard to imagine how this trivialization could have been done the other way around. I believe this very simple fact has been a powerful (if implicit) argument for the asymmetry of the tasks, as well as for the systematic down-grading of generation as a CL enterprise.

On the other hand, Chomsky's original transformational-generative (TG) grammar project (1966) served as an explicit argument for symmetry, though in a way that gave no comfort to any position in CL. The reason for this was that Chomsky always insisted that no procedural interpretation could be imposed on the operation of a TG: that it bound sentence strings to underlying representations statically, in much the same way that a function binds arguments to values without any assumptions about their direction. Functionality was Chomsky's own metaphor and it turned out to be, of course, incorrect.

Sentence strings and "underlying representations" may have a relationship that can be technically expressed as a relation, but it cannot be a *function* for the same reason that SQUARE-ROOT cannot, namely, that an argument like 4 is bound by the relation to the two values [plus and minus 2] and a function (as every child knows) can yield only a single value. The relationship between underlying representation and surface string is one-to-many in both directions (to speak in the way Chomsky wished to eradicate) and so cannot be functional unless one credits success to the efforts that Ross and others made twenty years ago to constrain or reinterpret transformations so that that relationship was more or less one to one. Another such attempt was that of Lakoff (1987), who tried to reinterpret the relation of "generation" (in the procedural, non-Chomskyan, sense of "down-the-page-towards-the-sentence-string") as one of deduction. That effort guaranteed asymmetry in an even stronger way than Lakoff could have needed in order to make a firm break with Chomsky's views, since implication deductions, whatever they are, are certainly not symmetrical. Had Lakoff been correct, no analysis could ever have been performed on Generative Semantics principles (which is to say, in the reverse direction from string to

representation). Even the current vogue for abductive notions could not alter that fact.

Much of this debate is long dead, however, because it became clear that Chomsky's position (i.e., that generation was non-directional and did not necessitate sentence synthesis) was motivated by antipathy toward computation and processes, which most linguists of today do not share. Also, the revival of phrase structure grammar methods during the last decade has made the questions of how TGs were best described, and whether or not they could be reversed to do analysis, a matter of less pressing interest.

Nevertheless, I believe that, historically speaking, Chomsky's insistence on the importance of abstract symmetry, taken together with the fact that real TGs at the time were in fact, and plainly, asymmetrical and directional, were powerful motivations for the resurgence of interest in phrase-structure grammars (PSGs). With the return of PSGs came abstract non-directional syntactic specification and well-known algorithms for running them either way (e.g., Kay 1984).

Then followed the arbitrarily arranged, yet fecund, marriage between revived PSGs and what one might call the "prolog movement." This dynamic duo held out the promise of abstract, declarative, symmetry between analysis and generation from the same grammar, as well as the possibility of real reversible processes, what one might call "running the prolog code backwards," i.e., either:

Provide The String To The Top-Level Predicate And Get The Structure

or

Provide The Structure And The Predicate Provides Its Own Argument, The Initial String.

This could not be a functional relationship for the same reasons as before, but it promised real magic and true symmetry between analysis and generation. Before very briefly posing the question, "was the magic real or mere sleight of hand?" let us pause in order to demonstrate why arriving at this point was significant by itself.

First of all, any demonstration of this symmetry helped the case of those who were actually interested in natural language generation. In other words, if the processes of analysis and generation were symmetrical or even identical, then generation would be as "interesting" as analysis (whatever that means), no matter how much greater the weight of research traditionally devoted to language analysis.

None of those considerations were relevant to those committed to natural language generation, but, from the point of view of the CL field as a whole, the recent growth of interest in generation needs some explanation. Let me put it this way: it needs explanation unless you believe that CL and AI are mainly driven by considerations of fashion.

At least, it needs more of an explanation than one given by a distinguished colleague who wrote last year that since "the problems of analysis are settled, the outstanding issues are now in generation." Since the assumption is false, this cannot serve as an explanation, however true its conclusion.

Let us now return to the prolog–PSG marriage, and ask whether it has yielded any real symmetry of processes (alias "reversible prolog code"). This question is not as straightforward as it appears to some. For instance, at NMSU–CRL, we have a multilingual machine translation program called ULTRA (1990) and, for the five languages it deals with (Chinese, Japanese, German, Spanish and English), we say we have "reversible code," meaning that by judicious predicate arrangement and avoidance of cuts, etc., the system is able to run a single chunk of prolog code for the analysis and generation of each language.

Moreover, our claim is not trivial, in that there is in fact a single set of syntactic rules for each language which can be used for both analysis and generation. In other words, the top level predicate does not simply hide a disjunction of two programs.

Nevertheless, the claim is still "cheap reversibility" insofar as it is only behavioral, specifying the minimum behavior required by the top-level predicate. There is certainly no claim that exactly the same set of predicates is evaluated in reverse order during analysis and generation of the same sentence, which would probably be impossible to achieve (even though the highest level sub-predicates in ULTRA do obey that reverse order condition). I mention this only to pose the more serious question: supposing one can have a certain amount of reversibility like that (and hence symmetry in the sense we started with), why should one want it, what benefit does it bring and what independent arguments are there in support of it? To say it has been a long-held human dream, like going to the Moon or climbing Mt. Everest, is an insufficient explanation and says no more than that we do it to prove that we can.

Arguments For and Against Symmetry

I think the abstract arguments for symmetry are:

1. the (half-forgotten) ''abstract symmetry'' argument inherited from Chomsky which I already mentioned;

2. a simple-minded use of Occam's razor, which forbids multiplying entities beyond necessity (although he said nothing about processes or modules or grammars).

Against the case for symmetry of generation and analysis are the following considerations:

1. that independent modules should be separated wherever possible. I know of *no serious* evidence for this, as an abstract matter, as opposed to programming practice, or some vague reference to Simon's (1969) ''decomposability'' notions for robustness or Jacobs' (1988) wish that modules should ''use as much shared knowledge as possible.'' This was also the sort of argument some transformationalists used in support of keeping syntax and semantics as separate modules, although the very same argument told against their own desire to hold analysis and generation together under a single description;

2. the psycholinguistic evidence suggests that analysis and generation are separate cognitive processes. One cannot get a single view on this from the literature, but we all know common-sense versions of it, such as the skills of a dialect speaker who can analyze Standard English, but hardly generate any. ''Passive vs. active competence'' is a well-attested notion in the field of language education and implies strictly separate cognitive skills.

Again, it seems obvious to many researchers that the choices required in analysis, which are largely of structural and lexical ambiguity, cannot be the same as the choices presented during generation because the latter are not mandatory. A good (if tired) example is a sentence containing the word ''bank,'' where the hearer has to decide, consciously or unconsciously, which meaning is intended. It is usually assumed that it is not necessary to do that in order to use (generate) the word ''bank'' appropriately in a sentence. I shall return to this later, but for now the traditional symmetry argument seems less persuasive after examination and the fact that some non-trivial reversibility of code, such as that one described above, is possible, tells against it.

Considerations of style, at the lowest level of choice, namely word paradigms, are similarly relevant. Consider the following:

*Roast fish (vs. meat)
*Rancid meat (vs. fats and oils)

These are pretty clear phenomena for native speakers of English (what I would call ''word preferences'') but ones whose violation does not impede their comprehensibility in the least.

In the ''roast fish'' case, one could re-express the asymmetry argument as follows: we understand it without difficulty but would not choose to generate it, preferring to say ''baked or broiled fish'' instead.

Do the analysis and generation activities for this phrase result from the same static information (what I am calling the word-preference of ''roast'')? It seems so. Are they done by the same processes in reverse? On the face of it, it seems they are not, because only analysis goes through the process of noting that ''roasting'' requires ''meat'' to follow it and that ''fish'' does not satisfy that requirement, though it has access to enough relevant properties of fish to allow the creation of a combination meaning.

But if speaking a language is to utter new and creative metaphors all the time, as many researchers assert, then we can also presume that a language generator must have access to the inverse of that very same process, since many metaphors have exactly the form of ''roast fish,'' e.g., ''rubber duck.'' If so, another apparent argument for asymmetry weakens in front of our very eyes. Nothing I am saying here calls into question the large-scale textual demonstrations by Church (1989) and others which show how such preferences are frequently violated in real text. These conventions are not overthrown by distribution any more than the standard generation of metaphors in the sentences of actual speakers overthrows the type-preferences they ''violate.''

Nothing said here *proves* that the same procedures are accessed in both direct directions. But the same issues arise in connection with plans as well as word-preferences. Jacobs (1990) has used considerations similar to these to argue that plans are not used in language generation, as many have believed when they assumed that a speaker first decides what to say and then plans how to do it. He uses examples like ''He answered the door,'' which is understood in preference violation terms as we described earlier (though Jacobs would probably not use quite that language), but which, he argues, is hard to explain in generation-as-planning terms, since it is hard to see why any planner would choose to generate that form at the word level. Jacobs's position is (like the final position of the present paper) basically a symmetricist

one, which sees no strong need for plans in either process.

Finally, we might look for further evidence for asymmetry by asking the question: is a "connectionist-based generation system" a contradiction in terms? Perhaps it should be, if it means training a system by feeding it hand-crafted sentence-pairs and representational structures. A connectionist system would lend credence to the symmetry case only if a single network could function for both purposes, but can that actually be done? One can *imagine* a network that yields strings for structures as well as structures for strings, but there remains a problem of how one would describe such training. Simmons (1990) has reported a connectionist yet symmetrical system, though it is not yet clear (to this author) whether or not it is of this form.

Is Semantic Parsing (SP) an Argument for Asymmetry?

Semantic parsing, you may recall, is a method claiming that text can be parsed to an appropriate representation without the use of an explicit and separate syntactic component. It was normally assumed (by Schank 1975 and others) that generation from a representation, however obtained, required the use of syntactical rules to generate the correct forms, even though a principal feature of SP was its claim to be the most appropriate method for analyzing (ubiquitous) ill-formed input without such rules. SP became reformed sinner when it came to generation. But is that assumption correct? Can we be sure that the so-called "arbitrary" choices made in generation are more or less arbitrary than those made in analysis?

Consider the following argument which demonstrates that the processes must be more or less symmetrical *even for SP:*

John loves Mary

In this (all-time favorite) example it can be argued that determining the identity of the agent and the patient is the same process for both analysis and generation. This argument is quite separate from an earlier one I made about lexical and structural ambiguity. Indeed, the present argument turns out to be none other than the traditional anti-SP argument that says that, if any system is able to distinguish the left-right order of symbols in a string, *then* the system has a syntax component. This argument is right about syntax only in Tarski's sense of abstract symbols, but not in the sense of a set of linguistics-style rules. SP proponents considered left-right order to be their own linguistic province as much as anyone else's.

The counter-examples to the "John loves Mary" argument for symmetry were other favorites like:

Economics, I like
Over the wall came a sturdy youth

In these examples, word order is less important to the interpretation than the fitting of entities to the preferences for the verbs. The first example would have the same meaning even if English did not mark the nominative case of the personal pronoun.

By the same token, the reason you cannot make

John loves Mary

or

Man eats chicken

mean their order inverses is not their syntax so much as the particular argument symmetry of these verbs (so that the inverses "make perfect sense"), which cannot be said of "like" and "come."

Now, SP might yield the symmetry of generation and analysis if a generation system based on coherence and best-fit slot filling were possible. In fact, on one occasion I designed a generation system for French (the Stanford MT project, 1973), that did have just that form. An SP (a preference semantics analysis of English) produced a "maximally coherent" structure of semantic objects from which a complex of corresponding French objects was produced. This was then "unwrapped" in such a way as to discard much of its content while separating out the most coherent subset that could be in a one-to-one relationship with a well-formed string of French words. On reconsidering that work, I realized that SP does not necessarily lead to a position in favor of asymmetry, as I had previously assumed. At that time, I argued that the structure at the intermediate representation level (in which French information replaced English) could itself be interpreted as a structure of gists, or messages, that the original conveyed. But nothing in the process, including whether the French output was an adequate equivalent of the English input, depended on that interpretation. Indeed, one could argue that a symmetrical SP (an SP followed by an inverted-SP generation) is compatible with a case in which there is no interpretable content at all.

This is the crucial point that brings us back to the second aspect of Wittgenstein's story mentioned earlier. As we can see, what matters there is not

the directionality of the counting, but our inability to imagine where the counting began. Classic natural language generation, on the other hand, starts from a structure which is simultaneously arbitrary and *interpretable*.

A traditional linguistics approach that utilizes a logical predicate structure as an underlying representation is closest to that ideal, while a connectionist system, where the lack of any internal representation is a virtue, is at the other extreme. The symmetrical SP I have described falls somewhere in the middle and could, in fact, go either way.

There has been much continued scepticism about such an interpretable intermediate representation both in AI and in philosophical thought related to it. Dennett, for example (1978), has remained sceptical about it. His original metaphor for language generation was that of a president's public relations officer who hands his leader a statement which is simply read aloud. Such a president may have no message he wants to convey; rather, he will simply say what "comes" to him.

A more succinct version is E. M. Forster's famous quip: "How do I know what I mean till I see what I say?" This takes the argument a step further and effectively concedes primacy to generation in the process of understanding. The philosopher and novelist both want to deny that there is any other code to be understood apart from (generated) natural language.

One can then raise the objection, what difference can all this scepticism about cognitive representations and our access to them possibly make? After all, AI is an abductive science and has never been deterred by any such lack of access to "human cognitive representations": normally it just invents such structures.

Nevertheless, one can retain a degree of healthy scepticism about them not only on the basis of lack of access, but rather on the logical nature of the structures classically preferred. Suppose that one felt sceptical about any internal representation (i.e. a message that represented what we "want to say" and from which we are able to generate) that was not in a real, natural, language. Dennett's public relations example is consistent with this point of view, although it is not normally used in support of it.

This supposition is also a form of Fodor's Language of Thought (LOT) hypothesis (1976), if the latter could be considered a natural language (NL) at all. Fodor has always been coy about revealing exactly what properties LOT may share with NL, though much of his writing implicitly claims that LOT is, in fact, a natural language (one which falls within a class defined by the parameters of Chomsky's Universal Grammar) that we simply do not have access to right now. It is obvious that, if the LOT hypothesis is true, then generation is, quite literally, a form of translation, albeit from a language we are unfamiliar with. A LOT could in principle be Italian or Aymara.

This is an important subject in itself but, in conclusion, I would argue that all this suggests that generation is not an independent computational subject like machine translation. In the latter case, one has true access to what one wants to say (the source language) and a firm sense of direction: one is familiar with both the source and target languages and the direction in which one is headed. Generation may then turn out to be a set of techniques that cannot be separated from a greater whole. But the same would hold true of analysis (as a separate subject) if we accept the earlier conclusion that the processes are fundamentally symmetrical. Yet everyone retains some nostalgia for asymmetry, certain that heading to some unknown destination is less disconcerting than coming from it.

The attempts I have made here to demonstrate that the two processes, of analysis and generation, are asymmetrical (as I would have wanted on my initial, unexamined, SP assumptions) have failed, and therefore to the credit of generation as an intellectual task even if not an independent one.

References

Appelt, D. 1989. Bidirectional Grammars and the Design of Natural Language Generation Systems. In Y. Wilks (ed.), *Theoretical Issues in Natural Language Processing*. Hillsdale, NJ: Erlbaum.

Chomsky, N. 1966. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Church, K., et al. 1989. Parsing, Word-Association and Typical Predicate Argument Relations. In M. Tomita (ed.), *Proceedings of the International Workshop on Parsing Technologies*. CMU.

Dennett, D. 1978. *Brainstorms*. Montgomery, VT: Bradford Books.

Farwell, D., and Y. Wilks. 1990. ULTRA: A Multilingual Machine Translator. Memoranda in Computer and Cognitive Science, MCCOY 202. Computing Research Laboratory, Las Cruces, NM.

Fodor, J. 1976. *The Language of Thought*. New York: Crowell.

Jacobs, P. 1988. Achieving Bidirectionality. *Proceedings COLING-88*, Budapest.

Jacobs, P. 1990. Why Text Planning Isn't Planning. In *Proc. NATO Advanced Workshop on Computational Theories of Communication*. Trentino, Italy.

Kay, M. 1984. Functional Unification Grammar. In *Proc. COLING-84*. Palo Alto, CA.

Russell, G., S. Warwick, and J. Carroll. 1990. Asymmetry in Parsing and Generating with Unification Grammars. In *Proc. Conference of the Assn. for Computational Linguistics*. Pittsburgh.

Schank, R. C. (ed.) 1975. *Conceptual Information Processing*. Amsterdam: North-Holland.

Simmons, R., and Y.-H. Yu. 1990. Training a Neural Network to Be a Context-Sensitive Grammar. In *Proceedings of the Fifth Rocky Mountain Conference on AI*. Las Cruces, NM.

Simon, H. 1969. The Architecture of Complexity. In *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Wilks, Y. 1973. The Stanford Machine Translation Project. In R. Rustin (ed.), *Natural Language Processing*. New York, NY: Algorithmics Press.

Winograd, T. 1972. *Understanding Natural Language*. Edinburgh: Edinburgh University Press.

Wittgenstein, L. 1956. *Remarks on the Foundations of Mathematics*. Oxford: Blackwell.

25
The Place of Heuristics in the Fulcrum Approach to Machine Translation

Paul L. Garvin

Theoretical Foundations

The theoretical conception on which the Fulcrum approach is based is the definitional model of language.[1] In this conception, the system of a language is considered to be, not a single hierarchy with a single set of levels ascending from phonology to semantics via syntax, but a multiple hierarchy structured in two dimensions, at least one of which in turn has three planes, with a separate set of levels proper to each of the planes.[2]

Language is viewed as a system of signs structured in two dimensions, those of the grammar and the lexicon. These two dimensions differ in terms of the purpose to which the signaling means of the language are put: the lexical dimension is defined as the system of reference to culturally recognized types of phenomena; the grammatical dimension is defined as the structure of discourse.[3]

The grammatical dimension of language is characterized by three planes, each with its own set of distinctions: the plane of structuring, characterized in all languages by two levels of structuring—those of phonemics and morphemics; the plane of integration, characterized in all languages by several levels of integration (the number of which varies from language to language); the plane of organization, characterized in all languages by two organizing principles—those of selection and arrangement.

All of these distinctions are defined by functional criteria:

1. The two levels of structuring differ in terms of the extent to which the units of each level participate in the sign function (meaning) of the language. The units of the phonemic level function primarily as differentiates of the sign function, the units of the morphemic level function as its carriers.

2. The levels of integration differ in terms of the order of complexity of the units that constitute them: they range from the level of minimal units, which is the lowest, to the level of the maximal fused units, which is the highest. Fused units are considered to be not mere sequences of units of a lower order, but to function as entities of their own order, with certain overall qualities above and beyond the mere sum of their constituents.

A correlate of the concept of fused units is the conception that the internal structure and the external functioning of a given unit are separate and potentially independent characteristics: units with the same internal structure may have different external functioning; units with different internal structure may have the same external functioning.

Units with the same internal structure are called identically constituted; units with the same external functioning are called functionally equivalent.

3. The two organizing principles on the plane of organization characterize different manners in which the signaling means of the language are employed: selection from an inventory versus arrangement in a sequence.

The three planes of the grammatical dimension of language are in a hierarchical relation to each other. The plane of structuring is defined by the most significant functional criterion and is therefore superordinate to the other two planes. Of the latter, the plane of integration is in turn superordinate to the plane of organization. Consequently, within each level of the plane of structuring a set of levels of integration can be defined, and within each level of integration of either level of structuring, the operation of both organizing principles can be discerned.

This conception of the structure of natural language is only an approximation: like all natural objects, natural language exhibits many indeterminacies and is more complex than any conceptualization of it can be.

One conspicuous instance of the intederminacies of natural languages is the perturbation of the covariance of form and meaning (which follows from the sign nature of language) by the well-known phenomena of homonymy and synonymy. Another instance is the lack of precision in the separateness of the levels

of language, as shown by the presence of some aspects of meaning (rather than mere differentiation) in certain phenomena usually assigned to the phonemic level of structuring (for instance, intonation, emphatic stress).

The complexity of natural language is apparent from the observation that in its overt manifestations (text, speech behavior, etc.) the different aspects (dimensions, planes, levels) of its underlying structure are not displayed separately but are closely intertwined, in the sense that each individual manifestation of the system displays all of its aspects together in a complex signal.

It is because of these indeterminacies and complexities that the model chosen for the conceptual representation of natural language is not quantitative, but qualitative. The model postulates only the general attributes of the object of study, but not the specific values and detailed manifestations of these attributes. These are to be ascertained by empirical means. Thus, the statement of the structure of a particular language is not considered a theory of this language, but rather a description within the frame of reference provided by a theory.[4]

In a linguistic description based on the definitional model, the various features of the model determine the organization of the description as follows:

1. The concept of the separateness of the two dimensions of language provides the justification for limiting the description to either of the two dimensions, and for keeping the grammar separate from the lexicon;

2. The concept of the levels of phonemics and morphemics on the plane of structuring provides the reason for differentiating the description of the phonemic pattern from that of the morphemic pattern, and to deal with their interrelations as a distinct aspect of the description;

3. The concept of the levels of integration provides the reason for organizing the description in terms of both minimal units and various orders of fused units, on both the phonemic and morphemic levels of structuring;

4. The concept of the potential independence of internal structure and external functioning provides the reason for differentiating these two aspects of linguistic units throughout the description;

5. The concept of the organizing principles of selection and arrangement on the plane of organization provides the reason for including in the description not only the inventories of units but also their distribution.

In the development of the Fulcrum approach, the primary concentration has not been on the further elaboration of the theoretical model of language, but on the design of a system appropriate to the task of translation, as well as the conduct of appropriate experimentation to test the adequacy of the system to the task. In the design of this system, the various features of the definitional model of language have served as guidelines but, by contrast with some other approaches to language data processing, the Fulcrum system is not intended to be a direct computer implementation of the underlying model. Rather, the function of the model is, from an operational point of view, to serve as a frame of reference for the design of the system, and from a theoretical point of view, to provide an explication and justification for the system.[5]

In this connection, it is important to note a basic difference between the application of the definitional model to linguistic description, and its application to the design of a machine translation system.

As was noted from the above, the organization of a linguistic description closely follows the hierarchic structure of the model. This is because, on the one hand, the model is considered a conceptual representation of the phenomenon of natural language in terms of its general properties, and on the other hand, a linguistic description presents the specific manifestation of these general properties in the case of a particular language.

In the design of the Fulcrum system, on the other hand, the properties of language as stipulated by the definitional model are taken into account in the order in which they are relevant to the process of translation. This order does not coincide with their organization within the model and the linguistic description.

Thus, the plane of organization, which ranks low in the hierarchy of planes of the grammatical dimension, is of primary significance in the theoretical interpretation of the translation process. The two organizing principles of selection and arrangement have been identified as the two basic components of the translation process since the early days of machine translation research (Garvin 1956).

Of at least equal importance is the plane of integration. The syntactic recognition routines of the translation algorithm are formulated in terms of the requirement of identifying the boundaries and functions of syntactic fused units (Garvin 1963b).

The plane of structuring applies to machine translation in the relatively obvious sense that the machine-readable input symbols (letters, spaces, etc.) belong to the graphemic level (which is functionally equivalent to the phonemic level of spoken language), while the units manipulated by the translation algorithm belong to the morphemic level (primarily words and syntactic fused units). The conversion from graphemic to morphemic units is accomplished by the dictionary lookup and by those subroutines of the translation algorithm which assign grammar codes (and with them morphemic status) to graphic elements not contained in the dictionary (such as symbols, missing words, etc.).

The two dimensions of language, which are kept separate in linguistic description, are taken into account together in the Fulcrum algorithm. The dictionary lookup is supplemented by special subroutines (such as the idiom and word combination routines) which allow the processing as single translation units of not only individual words, but also multiword lexical units. The syntactic recognition routines then treat these lexical units in the same way as syntactic units of similar structure that have been identified on the basis of purely grammatical criteria.

General Characteristics of the Fulcrum Approach

The Fulcrum approach differs from other approaches for automatic sentence structure determination primarily in the following respects:

1. The Fulcrum approach favors a bipartite, rather than a tripartite, organization of the parsing system.

2. The Fulcrum approach is characterized by two basic operational principles: (a) the concept of the fulcrum; (b) the pass method.

3. The Fulcrum approach aims at producing a single interpretation of each individual sentence, rather than at producing all conceivable interpretations.

Each of these characteristics will now be discussed further.

Bipartite Organization

A bipartite parsing system consists of two basic components: a dictionary with grammar codes (and other codes), and an algorithm which contains both the processing subroutines and the information required for processing. A tripartite parsing system consists of three basic components: a dictionary with grammar codes (and other codes), a processing algorithm, and a separate store of information (such as a table of grammar rules and other rule tables) which is called by the algorithm. The basic difference between these two types of system thus is that in a bipartite system the information required by the algorithm is written right into it, while in a tripartite system processor and information are kept separate.

Two types of advantages of the tripartite approach are usually cited by its proponents:

1. It separates the labor of the programmer who designs and maintains the processor from that of the linguist who designs and maintains the table of rules. The only thing they have to agree on is the format of the rules that the processor can accept. This minimizes the communication problem between linguist and programmer, since once these matters have been settled, the two portions of the program can be handled separately.

2. The same processor can be used with more than one table of rules. This means first of all that rules can be modified or changed without having to change the processor, provided of course that the format is maintained. This gives the linguist great freedom of experimentation with different types of rules. It also permits the use of the same processor for the parsing of more than one language, by simply substituting one table of rules for another.

These advantages apply particularly well to small experimental systems oriented towards linguistic research; for larger-scale experimentation, oriented towards the processing of randomly chosen bodies of text with the ultimate aim of designing an operational translation system, the advantages of a tripartite system are less clear-cut. This is why the Fulcrum approach favors a bipartite organization of the parsing system.[6]

The algorithm of a bipartite system is essentially not a ''parser'' of the type used in tripartite systems. It is instead a linguistic pattern recognition algorithm which, instead of matching portions of sentences against rules stored in a table, directs searches at the different portions of the sentence in order to identify its grammatical and lexical pattern. Thus, the essential characteristic of the algorithm is the sequencing of the searches, and in each search subroutine, only as much grammatical and lexical information is used as is appropriate to the particular search. The rules of the grammar and lexicon are in fact applied by the algorithm in a definite order, and a given rule is not even called unless the previous searches have led to a point where its application becomes necessary. This

means that the highly complex system of rules that makes up the real grammar and lexicon of a language is distributed over a correspondingly complex algorithm which applies the rules in terms of the ordering that the structure of the language requires.

The description of Russian which furnished the information included in the Fulcrum algorithm is based on the definitional model of language. It was developed using conventional Russian grammars and dictionaries as a starting point, verifying the reliability of the information, and adapting it to the requirements of the Fulcrum approach. In this process, it was found that many of the conventionally accepted statements about Russian grammar are not only inaccurate, but also that they are insufficient for purposes of automatic syntactic recognition. This is particularly true with respect to government, complementation, and mandatory co-occurrence relations.

Fulcra and Passes

A bipartite system stands or falls by the manner in which the problem of the sequencing of the searches within the algorithm has been solved. This is the key problem in developing the detailed structure of the algorithm.

The Fulcrum approach attempts to solve this problem by using two fundamental principles: the concept of the fulcrum and the pass method.

The concept of the fulcrum implies the use of key elements within the sentence (fulcra) as starting points for the searching through a sentence, which does not simply progress from word to word, but in fact 'skips' from fulcrum to fulcrum. It performs a little search sequence each time it has reached a fulcrum, and goes on to the next fulcrum when this particular search is completed.

The pass method means that not one, but several passes are made at every sentence, each pass designed to identify a particular set of grammatical conditions pertinent to the recognition process. Consequently, each pass has its own set of fulcra and its own search sequences. The pass method reflects the orderly progression in which the determination of the structure of the sentence is made: first, the sentence components are established, and finally the structure of the sentence as a whole is established. To each of these intermediate parsing objectives there corresponds, roughly, a pass or series of passes in the algorithm. The correspondence is not exact, because there are many ambiguities and irregularities interfering with the recognition process, and the design of the Fulcrum algorithm reflects these added complexities.

Single Interpretation of Each Sentence

Many automatic parsing systems are theory-oriented: their aim is to apply, verify, or otherwise deal with, a formal model of language, such as, for instance, a particular variety of phrase-structure grammars. One of the significant theoretical results of the use of such a parsing system is the determination of all the conceivable parsings that a given sentence is assigned by a particular grammar.[7]

The Fulcrum approach, on the other hand, is translation-oriented. Its aim is primarily to produce as correct a translation as possible. Clearly, for this purpose, the identification of all conceivable parsings of a given sentence is of no great interest. Rather, it is desirable for the algorithm to produce, at all times, if not the correct parsing, at least the most likely parsing of each sentence, to serve as the basis for its translation from Russian into English. In the earlier versions of the Fulcrum approach, this unique parsing was chosen deterministically on the basis of the contextual information available to it: for each set of conditions as identified by previous and current searches, the single possible—or most probable—interpretation was assigned to each syntactic and lexical configuration.

Thus, Russian clauses in which a nominal structure, ambiguously either nominative or accusative, both precedes and follows a predicate that agrees with either nominal structure, were interpreted by the algorithm on the basis of the highest probability in syntactic terms: the structure to the left of the predicate was interpreted as subject, that to the right of the predicate as object. The alternative interpretation (object-predicate-subject), although theoretically conceivable, was ignored. In the overwhelming majority of instances of course, this turns out to be the correct interpretation, as shown by the Russian one-clause sentence: Это предложение сохраняет нормальный порядок. which has only one reasonable interpretation: "This sentence preserves normal order."

There are a few structural configurations in which this probabilistic interpretation is not necessarily (or not at all) the correct one. First of all, there are some Russian clauses which, when used out of context, have only the one reasonable interpretation consisting of subject-predicate-object. But, because of their particular lexical structure, they require the alternative interpretation in certain contexts. So, for instance, the Russian one-clause sentence Автобусы заменили тролеибусы, would ordinarily be interpreted as: "Motor buses have replaced trolleybuses." But not so in the special context in which this sentence is pre-

ceded by У нас уже нет автобусов "We no longer have motor buses".[8] This context requires the alternative interpretation of object-predicate-subject: "Trolleybuses have replaced motor buses". (A stylistically better English translation would preserve order and replace the active predicate by a passive: "Motor buses have been replaced by trolleybuses".) There are, finally, a few Russian clauses which in any context have only the alternative interpretation (object-predicate-subject). The classical example of these constructions is Большой интерес представляет вопрос ... which, because of its particular lexical structure, can only be interpreted as object-predicate-subject: 'Of great interest is the question ...'.

The principle followed here is that, as the searching capability of the algorithm increases, the likelihood of erroneous choices decreases correspondingly. Thus, by increasing the lexical recognition capability of the algorithm, constructions of the last-mentioned type, in which lexical conditions override the effect of the syntactic configuration, can be identified and translated correctly. By increasing the range of contexts that the algorithm can search, constructions of the first-mentioned type, in which contextual factors override the effect of the syntactic configuration, can be identified and translated correctly. Clearly, the former recognition problem is much easier to resolve than the latter, since it requires only that special lexical meanings be taken into account, while the latter requires a form of "understanding" by the algorithm of the specific content of individual sentences.

Problems of the type just discussed are still within the capabilities of a deterministic recognition algorithm. There are, however, a number of identification problems of a different type which transcend the scope of a deterministic resolution capability and which require a heuristic approach to syntactic recognition. These will be discussed in the subsequent section.

The Need for Heuristics

The problems of the types treated in the preceding section do not require a revision of the basic design of the earlier versions of the Fulcrum algorithm. They do require access to more information of more kinds, but within the framework of the original pass method perhaps with an increased number of passes, or an improved overall layout of passes.

There are, however, a number of recognition problems for which the original deterministic design is inherently inadequate. These are the cases in which the correct resolution of a problem arising in a given pass requires the use of information that only a later pass can provide. From the standpoint of syntactic and lexical configuration, these are the instances in which the immediate context suggests the probability of a certain identification which, however, in the light of the total context of the sentence turns out to be incorrect.

The classical example of this type of configuration is the genitive singular/nominative-accusative plural ambiguity of nominals, the resolution of which as a genitive is suggested by an immediately preceding nominal structure. This identification, though correct in the majority of examples in Russian technical text, may turn out to be erroneous if other conditions in the broader context prevail; for instance, if a plural subject is required for the predicate of the clause and only the ambiguous nominal is an available candidate. This configuration is shown by the nominal задачи "of a task/tasks" in the clause В нашем плане задачи будут выполнеы "In our plan, the tasks will be fulfilled ..." Note that the resolution based on the immediate context is still likely to be the correct one in the majority of instances; it is the "usual" resolution which should be overridden only under "special" conditions.

One treatment of the type of problem illustrated by the above example would be for the algorithm to record both possible interpretations of the ambiguous form early in the program, and make the selection later when the information from the broader context has also become available. This solution would, however, fail to take into account the characteristic feature of this type of configuration, which is that the two possible resolutions of the syntactic ambiguity are not equally probable: in the majority of occurrences, a correct identification can be based on the immediate context, and the broader context has to be resorted to only under special conditions. This requires a method of resolution which will accept an identification based on the immediate context, will let it stand in the majority of cases, but will have the capability for revising this decision in all those cases in which the special conditions apply which call for an identification in terms of the broader context. Such a method of resolution is heuristic in nature; it is discussed in detail in the subsequent sections.

Heuristic Principles

The Fulcrum approach has borrowed the concept of heuristics from its applications in artificial intelligence research.

As is well known, the concept of heuristics is related to problem solving. This is how most students of artificial intelligence speak of it. According to M. Minsky (Feigenbaum and Feldman 1963:407), "The adjective 'heuristic', as used here and widely in the literature, means *related to improving problem-solving performance*, as a noun, it is also used in regard to any method or trick used to improve the efficiency of a problem-solving system." G. Pask (1964:168) speaks of "... a set of 'heuristics' ... or broad rules and suggestions for problem solution ...".

One characteristic of heuristics is that it is "provisional and plausible" (H. Gelernter in Feigenbaum and Feldman, 1963:135). Another more important characteristic is that they are "processes ... which generally contribute to a result but whose effects are not 'guaranteed'" (Newell and Simon 1963:390).

The major advantage of heuristic principles is considered to be that they "contribute, on the average, to reduction of search in problem-solving activity" (F. M. Tonge in Feigenbaum and Feldman 1963:172). Thus, "... a heuristic procedure substitutes the effort reduction of its shortcuts for the guaranteed optimal solution of an exhaustive method ..." (ibid. 173).

Theorists of heuristics often speak of heuristic processes. The mathematician G. Polya, who is often cited as an authority on heuristics by students of artificial intelligence, defines modern heuristics as the study of "the process of solving problems" (1957:129). He links the use of heuristics to plausible reasoning, as applied in the "heuristic syllogism", which he differentiates from the demonstrative reasoning of logic (ibid. 186–190). Others emphasize the methodological aspects of heuristics. Thus, E. A. Guillemin (1931:10) speaks of "... a method of solution ... which is used almost exclusively by physicists and engineers. This method is nothing more than judicious guessing. The elegant title by which this method is known is the heuristic method."

All of the above-noted aspects of heuristics have to do with the general functional characteristics of heuristic processes or methods. Clearly, they all are in some way pertinent to syntactic resolution in general and the Fulcrum approach in particular. We are dealing with a form of problem solving; the solutions may have to be provisional and plausible rather than definitive, and they are certainly not guaranteed; the Fulcrum approach, at least, has as one of its major aims the reduction of the number of required searches; certainly, all forms of syntactic resolution are based on plausible rather than demonstrative reasoning, and are in essence well-organized judicious guesses.

In view of all this, it might not be unreasonable to refer to all syntactic recognition procedures as recognition heuristics. The reason this has not been done is because in the Fulcrum approach a somewhat more specific and restricted definition of heuristics has been used than that implicit in the aspects listed so far.

Such a more specific definition is based on the design characteristics of a heuristic program, rather than on the general purpose of the heuristic approach. While these design characteristics are not explicitly stated in the literature, they can be extrapolated from an examination of the use of heuristics in artificial intelligence (cf. several of the articles in Feigenbaum and Feldman 1963). In essence, a heuristic program consists of an alternation of trials and evaluations based on a clearly defined strategy. The strategy is that of a problem solver, the trials are the "judicious guesses" (see above) which characterize the heuristic method, and the evaluation of the trials is based on criteria of goal attainment derived from a definition of the problem.[9]

Usually a heuristic program and an algorithm are considered two alternative ways to approaching a problem. Thus, A. Newell, J. C. Shaw, and H. A. Simon note (Feigenbaum and Feldman 1963:114) that there may be "both algorithms and heuristics as alternatives for solving the same problem." In the Fulcrum approach, on the other hand, heuristics is not used as an alternative to an algorithm. Rather, the two are combined in the same program: the Fulcrum algorithm contains certain heuristic portions designed for the resolution of only those identification problems that do not lend themselves to a straightforward algorithmic treatment. This means that the Fulcrum algorithm, in addition to the heuristic trial and evaluation components, must also contain provisions for identifying those sets of conditions under which heuristic resolution is required.

These design features of the heuristic portions of the Fulcrum algorithm will be discussed in the subsequent section.

Design of the Heuristic Portions of the Fulcrum Algorithm

As has been noted in the preceding section, the design of the heuristic aspects of the Fulcrum algorithm is not identical with that of an independent heuristic program. Rather, the need to adapt the heuristic design principles to the requirements of the Fulcrum approach has led to the development of a design quite specific to this particular purpose.

The most typical feature of this design has already been mentioned, namely, the overall characteristic that the heuristic is, as it were, embedded in an algorithm. Thus, the executive routine of the heuristic, which carries out the "guessing" strategy by calling the trial and evaluation routines, in fact constitutes a bridge between the deterministic main portion of the algorithm and the heuristic portion. It operates on the basis of a capability of the deterministic main portion of the algorithm for recognizing when to call the heuristic portion. This capability is one for recognizing the circumstance, already noted previously, that for a given ambiguously interpretable form the conditions present in the immediate context do not guarantee a correct identification. Once this recognition has been effected, the Fulcrum algorithm makes the transition from the deterministic main portion to the heuristic portion and acts as the executive routine of the heuristic.

The remaining aspects of the heuristic portion of the Fulcrum algorithm, namely, those dealing with the conduct of the trials and evaluations, likewise differ significantly in their design from an independent heuristic program.

An independent heuristic program, such as those used for game-playing or theorem-proving (see Feigenbaum and Feldman 1963), carries out more than one trial every time it "considers" a particular move or other operation. By contrast, the heuristic portion of the Fulcrum algorithm conducts only one trial each time it is called, or more specifically, it carries out a particular single syntactic identification in the form of a trial, subject to later revision. The question asked in an independent heuristic thus is, which of several trials (if any) is successful? The question asked by the heuristic portion of the Fulcrum algorithm is, is this particular trial successful?

In an independent heuristic, evaluation takes place immediately after each given set of trials has been completed. In the heuristic portion of the Fulcrum algorithm, the evaluation of a given trial identification does not take place until later in the program. This is because, as was repeatedly noted before, the trial identification is based on the broader context, and the Fulcrum algorithm deals with the immediate context significantly earlier in the program than with the broader context.

As in any heuristic, so in the heuristic portion of the Fulcrum algorithm, the essential subject-matter question concerns the factors on which the trials and evaluations are based.

In the heuristic syntax, the trials are based on probability: as has already been noted, a given trial identification is always made on the basis of the most likely solution suggested by the immediate context. It must be stressed that this likelihood is determined impressionistically on the basis of available knowledge of Russian grammar; it is not considered necessary to have recourse to a formal probability calculus. The evaluations are based primarily on the mandatoriness of certain syntactic relations within the broader context: if the broader context requires that a certain syntactic function (such as that of subject) be filled, and this condition can be met only by revising a previous trial identification, then this requirement constitutes the evaluation criterion on the basis of which the original trial is rejected and an alternative solution is substituted for it.

The heuristic portion of the Fulcrum algorithm operates in the following manner. Whenever the recognition routines identify a set of conditions under which a trial identification is made, a record of this trial is written (a heuristic "flag" is "set"). When later in the program the broader context requires a mandatory syntactic component for which no suitable candidate is present, the algorithm "looks for" a heuristic flag. If it finds a flag, the trial identification is judged a failure on the basis of the newly encountered conditions of mandatoriness, and the alternative identification is chosen in its stead, in order to satisfy this condition of mandatoriness.

As can be inferred from the above, the use of heuristics in syntax presupposes the inclusion in the grammar code of the Fulcrum system of all those indications that are essential to the operation of the heuristic portion of the algorithm. In particular, this means including information about mandatoriness of syntactic relations where this is not implicit in the word class of the dictionary entry. Thus, for every attributive (adjective or adjectival pronoun), a head is mandatory and hence no special mandatoriness notation is required in the grammar code. In the case of predicatives, on the other hand, a subject or object may be either optional or mandatory, and hence a mandatoriness notation in the grammar code is necessary.

Specific examples of heuristic ambiguity resolution in the Fulcrum algorithm are discussed in the subsequent section.

Application of Heuristics to Particular Syntactic Resolution Problems

Two areas of syntactic resolution will be discussed to illustrate the application of the heuristic portion of the Fulcrum algorithm. These are the syntactic

interpretation of genitive nominal blocks and the resolution of predicative-adverb homographs (word-class ambiguities of the type *SICI10*). Genitive nominal blocks here include both those that are unambiguously genitives and those that are ambiguously genitives. The latter are nominal blocks which in addition to the genitive function have other case functions, requiring the resolution of the case ambiguity in addition to other aspects of syntactic identification.

Genitive Nominal Blocks

The cases of interest here are those for which the immediate context suggests that the (unambiguously or ambiguously) genitive block functions as an adnominal genitive complement. This resolution may be overridden by conditions in the broader context which the heuristic capability of the program recognizes.

Thus, the ambiguous genitive полета "(of) flight" in the immediate context время полета "time (of) flight" will be identified as the adnominal genitive complement. However, the broader context may require that this genitive form be interpreted as the genitive of reference of a negative predicate, as when the above example is expanded to read: В это время лопета не было "at this time there was no flight." The heuristic capability of the program will then carry out the required revision of identification.

Other types of conditions in the broader context which may require heuristic revision are:

1. Genitive nominal block is required as head of a (governing) modifier;

2. Genitive nominal block is required as subject of a predicate;

3. Genitive nominal block is needed as object of predicate;

4. Genitive nominal block is required as genitive of subject or object of deverbative noun.

Note that in each of the above cases, a relation in the broader context (head of modifier, subject of clause, etc.) is considered mandatory. In order to comply with this condition of mandatoriness, the previous identification based on the immediate context is overridden, and an identification which satisfies the mandatory relation in the broader context is substituted.

The types of conditions listed above are illustrated by the following examples.

(1) Выполненные бригадой работы ...

The immediate context here suggests the trial identification of the ambiguously genitive noun работы "(of) work(s)" as the adnominal genitive complement to бригадой "(by) the brigade", to read бригадой работы "(by) the work brigade". The broader context, however, requires that a head be assigned to the nominative/accusative plural governing modifier (past passive participle) выполненные "performed", and the ambiguously genitive noun работы (which can also function as nominative/accusative plural) is the only available candidate. Consequently, the trial identification as genitive adnominal complement is rejected, and replaced by a definitive identification as head to the governing modifier. The sentence fragment is then interpreted correctly as reading 'work performed by the brigade'.

(2) В эксперименте цели будут выполнены ...

The immediate context here again suggests the trial identification of the ambiguously genitive noun цели "(of/to/by) goal(s)" as the adnominal genitive complement to эксперименте "experiment". The broader context, however, requires that a subject be assigned to the plural predicate будут выполнены "will be fulfilled", and the ambiguously genitive noun цели (which can also function as nominative/accusative plural) is the only available candidate. Consequently, the trial identification as adnominal genitive complement is rejected and replaced by the definitive identification as subject. The sentence fragment is then interpreted correctly as reading "In the experiment the goals will be fulfilled ...".

(3) Данный метод результата не дает.

The immediate context suggests the trial identification of the unambiguously genitive noun результата "(of) result" as the adnominal genitive complement to данный метод "(the) given method". The broader context, however, requires that an object in the genitive be assigned to the negative predicate не дает "does not give", and the unambiguously genitive noun результата is the only available candidate. Consequently, the trial identification as adnominal genitive complement is rejected and replaced by a definitive identification as object. The sentence is then interpreted correctly as 'The given method gives no result'.

(4) ... определение с максимальной точностыо формы диаграммы ...

Again, the immediate context suggests the trial identification of the ambiguously genitive noun формы "(of) form(s)" as the adnominal genitive

complement to точностыо "(by) accuracy". However, the broader context requires that a genitive of object be assigned to the deverbative noun определение "determination", and the ambiguously genitive noun формы is the only available candidate. Consequently, the trial identification is rejected and replaced by a definitive identification as genitive of object. The sentence fragment is then interpreted correctly as reading "The determination of the form of the diagram with maximum accuracy".

Predicative-Adverb Homographs

The cases of interest here are those for which the immediate context suggests that the homograph functions as an adverb. This resolution may be overridden by mandatory conditions in the broader context which the heuristic capability of the program recognizes.

Thus, the homograph понятно "is understandable/understandably" will be identified as an adverb in the immediate context понятно высказанное "understandably voiced". However, the broader context may require that this homograph be interpreted as a predicative, as when the above example is expanded to read: Нам понятно высказанное И. П. Павловым убеждение. что ... "We understand the conviction voiced by I. P. Pavlov, that ... (lit.: the conviction ... is understandable to us)". The heuristic capability of the program will then carry out the required revision of identification.

The mandatory condition in the broader context here is, of course, that a clause should have a predicate whenever any candidate at all is available. Since the neuter nominative nominal block высказанное И. П. Павловым убеждение qualifies as subject, and the nominal block нам qualifies as the appropriate dative object, the homograph reinterpreted as a neuter predicative will meet both the condition of agreeing with the subject and the condition of governing the object, thus providing the clause with the needed predicate.

Implementation of Heuristic Syntax

The essential characteristics of heuristic syntax as applied in the Fulcrum approach can be summed up as follows:

1. The heuristic portion of the Fulcrum algorithm is called whenever there is a possibility that a given identification made on the basis of the immediate context may have to be revised on the basis of information provided by the broader context.

2. The conditions requiring the use of heuristics are recognized by the deterministic portion of the Fulcrum algorithm.

3. The mechanism for calling the heuristic syntax consists in the writing of a record (setting a "flag") in the sentence image which the program produces, indicating that a given identification has been made on a trial basis and is subject to heuristic revision.

4. The evaluation criteria for the revision of trial identifications consist in various conditions of mandatoriness of occurrence of certain syntactic components. These conditions are recorded in the grammar codes of the dictionary entries which the Fulcrum algorithm manipulates. Some of these conditions are contained in the grammar codes by implication: thus, the word class code notation "modifier" implies the requirement of a head to which this modifier is to be assigned. Other conditions must be noted explicitly in the grammar code, for instance, the mandatoriness of subjects or objects for certain predicatives, or the mandatoriness of genitives of subject or object for certain deverbative nouns.

5. The mechanism for applying a heuristic revision to a trial identification consists of the following:

a. The program first notes the absence of a mandatory syntactic element by acting upon the requirements implicit in the grammar code, or by reading the specific mandatoriness notation.

b. The program now tests for the presence of heuristic decision records ("flags") in the sentence image and checks whether the recorded element is a suitable candidate for the missing syntactic component.

c. If these tests are positive, the trial identification is revised and a definitive identification is substituted for it.

As can be noted, the apparatus for the heuristic syntax consists primarily of a capability for recognizing the need for heuristics, suitable notations in the grammar code to allow the heuristic evaluation of trial identifications, and a mechanism for writing and reading heuristic records in the sentence image, on the basis of which the revision of trial identifications can take place.

Notes

1. This paper is a revised version of Progress Report No. 14 under Contract NSF-C372, 'Computer-aided research in machine translation', with the National Science Foundation.

2. For a detailed discussion of an earlier formulation, see Garvin 1963a. For a more recent, but more concise discussion, see Garvin 1968.

3. For a detailed discussion of the two dimensions, see Mathiot 1967.

4. The classical statement of the opposite view is found in Chomsky 1957:49: "A grammar of the language L is essentially a theory of L."

5. For a different conception of the role of the model in a machine translation system, see Lamb 1965.

6. For a more detailed discussion of the reasons for this preference, see Garvin 1966.

7. Cf. Kumo 1965:453: "A predictive analyzer produces for a given sentence all possible syntactic interpretations compatible with the current version of the predictive grammar."

8. I am indebted to A. Isacenko for this example.

9. For a more detailed discussion of this view of heuristics, see Garvin 1964: 80–85.

References

Chomsky, N. 1967. *Syntactic Structures*. The Hague: Mouton.

Feigenbaum, E. A., and J. Feldman, eds. 1963. *Computers and Thought*. New York: McGraw-Hill.

Garvin, P. L. 1956. Some Linguistic Problems in Machine Translation. In *For Roman Jakobson*. The Hague: Mouton, pp. 180–186.

Garvin, P. L. 1963a. The Definitional Model of Language. *Natural Language and the Computer*, ed. Paul L. Garvin. New York: McGraw-Hill, pp. 3–22.

Garvin, P. L. 1963b. Syntax in Machine Translation. *Natural Language and the Computer*, ed. Paul L. Garvin. New York: McGraw-Hill, pp. 223–232.

Garvin, P. L. 1964. Automatic Linguistic Analysis—A Heuristic Problem. *On Linguistic Method*. The Hague: Mouton, pp. 78–97.

Garvin, P. L. 1966. Some Comments on Algorithm and Grammar in the Automatic Parsing of Natural Languages. *Mechanical Translation*, 9:2–3.

Garvin, P. L. 1967. The Fulcrum syntactic analyzer for Russian. *Preprints for 2ème Conférence Internationale sur le Traîtement Automatique des Langues. Grenoble, 23–25 août 1967*. Paper No. 5.

Garvin, P. L. 1968. The Role of Function in Linguistic Theory. *Proc. of the X Internal. Congress of Linguists, Bucharest*.

Guillemin, E. A. 1931. *Communication Networks, Vol. 1*. New York: Wiley.

Kuno, S. 1965. The Predictive Analyzer and a Path Elimination Technique. *Communications of the ACM*, 8:453–462. Reprinted in David G. Hays, ed., *Readings in Automatic Language Processing*. New York: Elsevier, 1966, pp. 83–106.

Lamb, S. M. 1965. The Nature of the Machine Translation Problem. *Journal of Verbal Learning and Verbal Behavior*, 4:196–211.

Mathiot, M. 1967. The Place of the Dictionary in Linguistic Description. *Language*, 43:3.

Newell, A., and H. A. Simon. 1963. Computers in Psychology. *Handbook of Mathematical Psychology*, ed. R. Duncan Luce et al. Vol. 1. New York: Wiley, pp. 361–428.

Pask, G. 1964. A Discussion of Artificial Intelligence and Self-organization. *Advances in Computers*, ed. Franz L. Alt and Morris Rubinoff. Vol. 5. New York: Academic Press, pp. 109–226.

Polya, G. 1957. *How to Solve It*. Garden City, NY: Doubleday.

26
Computer Aided Translation: A Business Viewpoint

John S. G. Elliston

Before one starts to look for a particular solution, it is necessary to define the precise needs of the problem. Such is the case with our company (Customer and Service Education Ltd.); the solution we are pursuing is tailored to the specific communication needs we have identified and it may well not be the most effective direction for another company. In order to understand why we have chosen our particular path, it is helpful to explain briefly the company environment.

The Environment

Xerox operates in more than 36 countries spread throughout the world. The task of our Technical Service function is to install and maintain our products, both rented and sold, in each of these countries. Although the size of operation differs considerably between countries, the individual functional support need, in terms of technical data for our Service representatives, is virtually identical. The data is provided in the form of Service Documentation.

This documentation is vital for the field service organizations to be able to do their job. The documentation provides the Service Representative with all the technical data that he needs. It comprises maintenance procedures, technical data, diagnostic procedures and spare parts lists. This type of information must be provided for all products and all configurations. To provide all of this data, we go through the following processes: documentation development, validation, translation, production, distribution and maintenance. An operation of this type is complex enough, the pressure related to accuracy and timeliness for individual locations just adds to the difficulty. Materials are developed prior to launch and "in-field" validation tests in English may well be running concurrently with several translation programmes, in order to enable the staggered national launch needs to be met.

Of the 36 countries mentioned, only seven have English as their first language. Even within this group there are sufficient differences within the languages to cause some misunderstanding. A further 14 countries are obliged to use English text documentation largely because of economics related to the scale of operation. Within this group the ability to speak English varies from very high to very low. The remaining 15 countries require that all documentation is translated before it can be used in their field environment.

The whole operation is critical and costly and any inefficiency can quickly escalate costs. Consequently, the process needs to be subjected to tight controls.

The Problem

The problem can be broadly expressed under three headings:

- Costs.
- Timeliness or lapse time.
- Clarity of communication.

Costs

The demand on our translation resource grows every year. This demand is related to our increasing product range, refinements to existing products and the normal on-going need to maintain existing documentation. An additional factor is the legal demands placed upon a multinational operation to translate to meet legal requirements. One obvious answer is to increase our resource to handle the growing load. Unfortunately, increasing the translation resource increases our cost base and makes us less competitive. The solution we need must be found in productivity, i.e., using the resources we already have, more efficiently.

Timeliness of Lapse Time

Service documentation is developed by our headquarters function either in the U.S. or U.K. In either case, it will be originated in English. On average, it will be between three to four months after the first English version has been validated before a translated manual will be available. (The precise figure will depend upon the complexity of the product

and translation workload and prioritization.) This lapse time reduces the possibility to field-test products in non-English speaking environments and consequently, puts a heavy burden on the English-speaking companies, who must now do the majority of the field tests. The question of validation in non-English-speaking markets is further compromised by localized translation. Manual translation will inevitably be tempered with experience and interpretation. This means one is no longer testing the original, therefore results obtained are invalid. This is even more of a problem if the "subjective", or the "interpretative" translation actually improves on the original English version. It is one thing to identify a documentation fault and relate it to either an origination or translation error; it is quite another thing when a problem is resolved by the translation. In the latter case, the translation passes the test and the original English version gets printed complete with fault.

Timeliness is also a key feature of our documentation corrections and update system. At present, extensive delays can result before translated data is available to the field. Again, the reasons are the same, the complexity of the task and prioritization.

Clarity of Communication

The two major factors that contribute to ambiguity within our multinational environment are:

- ambiguity—text must be written in a clear manner.
- vocabulary—text should only contain those words that are known to be in the end users' vocabulary.

A commonly expressed opinion is that if a group of 50 translators were given the same sentence to translate, they would produce 50 different versions. A computer given the same sentence will only give one translation. How can we assume that the one output from the computer is the right translation? I believe that the question directs our attention to the wrong place. The real problem is in the fact that the original sentence was capable of 50 different interpretations. To the producer of Service Documentation, this is frightening. If one sentence is open to so much interpretation, what chance does a Service Representative have when one realizes the permutations of a complete book? Obviously, the first problem to tackle is the generation of source material.

Our experience to date has shown that it is extremely difficult to define clarity sufficiently objectively to ensure an author writes clearly. Each writer has his own personalized style. Simply using good grammatical English does not in itself eliminate ambiguity. If a writer has written a piece of text that conforms to grammatical rules, the question as to whether it is ambiguous usually results in, at best, subjective discussion and, at worst, emotive argument.

Secondly, with present techniques, it is not possible to ensure that authors only use those words within the vocabulary of the target population. Our target population spreads across 36 countries and ranges from 18 to 50 years of age.

Added to this situation, we find that all too often words or phrases have a different meaning when placed in a different context or worse still, in another cultural environment. Recently, whilst visiting the United States, I purchased a coffee from a secretary who looked after the departmental percolator. The price for the coffee was a very modest 10 cents. When I learned of the low cost, I mentioned that it was very cheap. This comment was followed by a rather obvious silence. My colleague later pointed out that it would have been better to suggest that the coffee was "inexpensive". The word "cheap" in the U.S. is usually used in a derogatory sense. Thus, my rather innocent comment was taken as a criticism of the coffee and could have resulted in my having to find an alternative source of coffee.

It is perhaps this type of apparently insignificant interpretation that can so easily result in misunderstanding, or even offense. This is especially risky where we tread the often delicate path of operating across national practices and customs.

The Need

The need is relatively straightforward. Our company needs a means of communicating technical data, instructions and information to our worldwide Field Service. The method chosen must be acceptable in the business sense, that is the costs incurred must be less than the benefits gained. It must be capable of providing the output communication when and where it is required. Thirdly, it must ensure that the end user can retrieve data accurately and quickly.

Finally, throughout the complete cycle from generation of source language text to translation into target language text, the needs of three categories of end user must be met.

- Personnel whose first language is English.
- Personnel who are obliged to use English but whose first language is not English. (This particular requirement adds a third dimension to the discussion

of manual vs. machine translation, as it automatically forces one to look more closely at the source language.)

• Personnel (or machines) who are required to translate the English text into the target language. (These people differ from category two insofar as they will not have the same depth of technical knowledge and understanding.)

The Solution

There was, and still is, no instant or obvious solution. The path we have followed has taken us through several potential solutions, each in turn being discarded until we have arrived at our present status.

Perhaps the first approach that we looked at, was one based on the "Caterpillar English" concept. At its simplest it is a limited vocabulary with each word being carefully defined. The target population is then taught to recognize the words rather as one would recognize a symbol, then associate it with the defined meaning. The end user is not taught to pronounce the words, just to recognize them, thus he does not actually learn the source language. This method has been successful in many areas, but did not fit our particular situation. It would be true to say that we rejected the system more on social grounds than on the basis of any real scientific testing. A new company setting up its operation may well find the system workable, although legislation in some countries might make even that difficult. In our situation, we were dealing with well established operating companies who already translated material to a high standard and a Field Service force, used to having their support documentation in their own language. To switch to a limited English language was seen as a retrograde step and totally unacceptable.

A second solution considered, was the use of a "Command English." This looked a far more likely solution as one could fairly accurately prepare translation for standard command sentences. This would achieve two things. Firstly, a guarantee that the translation is accepted in advance and secondly, a machine can be used to speed up the process. The difficulty that was encountered in this attempt was the constraints placed upon the source language writer. Much of the Service information can be expressed in the directive manner of command English. The problem starts to show when one writes "descriptive statements" or "test objectives" or even statements relating to judgements. In addition, the potential for developing the Command language for use in the areas of training and customer documentation seems almost zero.

During the period these approaches were being investigated, we also examined some of the claims at that time for existing computer translation systems. These systems by and large claimed to offer unrestricted input translation and seemed promising. Regrettably, these claims seldom lived up to the test and the systems tended to be extremely expensive in development and post-edit costs per language.

The system that we are currently using to develop our total translation process is SYSTRAN. Initially, we did some research with uncontrolled input text which resulted in unacceptable output in terms of the post-edit effort required.

The dilemma at this stage was that if one used a totally free form of input, the computer translation output required a massive post-edit. Conversely, if the source language was written to permit computer aided translation it became unacceptably restrictive to the author or originator. An additional problem with this tightly controlled input is the acceptance of the user of source language material.

The large post-edit task was unacceptable in terms of both cost and job satisfaction. The amount of post-edit was such that it took almost as long as it would have taken to translate the whole exercise manually. The morale of a translator in this mode of operation is low. After all, the job is reduced to trying to understand and correct a rather badly written document.

By now it was clear that computer aided translation was achievable but its acceptability was related to the balance between the control of the source language input and the degree of post-edit required of the target language output. On the one side, if the constraints placed on the originator are too severe the increased load would cancel the productivity benefit of the system. In addition, one runs into the real danger of author motivation. On the other side, if one relaxes the input control on the source text translation too much, the post-edit function grows to the point that machine productivity is wasted and a similar motivation problem exists, this time for the translator.

The input controls that we have placed upon text origination falls in two main categories:

• Vocabulary—It became necessary to ensure that misunderstanding or ambiguity did not arise out of the use of a particular word, or because of the context in which the word was placed.

• Writing Rules—Once again, to reduce ambiguity in the source text it was necessary to determine rules to define the required size and construction of sentences, etc.

The vocabulary was developed by combining the work initially done in our U.S. and U.K. based locations. For example, in the U.K. location we had developed a vocabulary from ILSAM (International Language for Service and Maintenance). This vocabulary became known as RX Customized Vocabulary. The vocabulary was compared to one that was developed in our U.S. location and from the two sets, we developed our present lists, now known as MCE (Multinational Customized English). This vocabulary is made up of several sub-groups.

Firstly, the basic core group vocabulary consisting of approximately 1000 words. This group forms the basic communication word list. The other groups are to permit the specialist communication within our specific company environment. They fall under the categories of copy quality terminology, publications terminology, abbreviations, weights, measures, etc. In all, this provides a total vocabulary of under 3000 words.

The next step in the process was to get each target language user to identify their own language equivalents for each word in the MCE vocabulary. As anyone who has tried will know, selecting one foreign language word for one English word is a tough proposition. The important factor is not to simply look for a word for word equivalent, but define one and only one meaning for each English word and then find the target language word or phrase to relate to that precise meaning. For example, the word "replace" is often used to request two quite different actions, e.g.,

• Remove part A and replace it with part B.

• Remove part A, adjust part B and then replace part A.

In the first case, we are using the word to mean "exchange" and in the second case to mean "put back." This usually gives little problem to experienced English speaking staff, but does cause problems for those who use English text, but whose first language is not English. It also gives problems to the computer.

Again, for the sake of clarity, each word was defined as a specific part of speech and, if possible, never in more than one category, i.e., "switch" as a noun and not as a verb. Unfortunately, this was not always possible.

At this point of time in the development cycle, we are gradually being forced to face a simple truth. The computer refuses to understand unless we write clearly and simply. This should not be a lot to ask, in fact it is really what our Service Representatives have always required. Seen in these terms, the project seems reasonable. If we are writing service support documentation with a vocabulary that people are not familiar with, then we are clearly not doing our job effectively.

The same correlation is found between the needs of the end user and computer in terms of writing rules. If sentences are written simply and kept short, then the target population is satisfied. For example, an English technician may have no problem with the following statement:

"Loosen main motor and drive shaft and slide back until touching back plate."

This statement demonstrates at least two problems. Firstly, the sentence is too complex. It needs to be written in several short sentences. Secondly, in an attempt to reduce the amount of text the technician must read, we tend to leave out the definite article. Again, anybody familiar or trained on the subject and who speaks English fluently will probably have no problem. The computer unfortunately has neither of these two advantages.

Imagine reading a telegram (usually written in abbreviated form to save costs), "SHIP SINKS." Does it mean "THE SHIP SINKS" or "SHIP THE SINKS"? The difference in meaning by simply moving the position of the definite article is enormous. To overcome these difficulties, the original statement should be written as follows:

"Loosen the main motor. Loosen the drive shaft.
Slide both parts until they touch the back plate".

Summing up, the input to the computer is controlled to the extent that it must be written within the vocabulary of the computer and written in simple short sentences.

One concern that was originally felt by the English speaking service representatives, is that the text that they would be issued with would be written in a form of pidgin English. The example in figure 26.1 shows this is clearly not the case. Our field test indicated that of the order of 90% of our U.K. sample found no difference between the ordinary text and the customized text, in terms of usability. Our tests in Sweden, where English is the second language, indicated a 70%

CHECKS AND ADJUSTMENTS

WARNING: MALADJUSTMENT OF PLATEN COVER INTERLOCK SWITCH S15 AND INCORRECT OPERATION OF THE MERCURY SWITCH S2 (16.3), CAN EXPOSE OPERATOR TO EXPOSURE FLASH. ENSURE BOTH SWITCHES ARE OPERATING CORRECTLY.

1. (B5, B6) Loosen screws (G) front and rear.

1. Loosen front screws (G). Remove switch S3. Loosen spacers and screw, that secure rear hinge.
2. Put (B5) 600 T 91030, (B6 onward) 600 T 91197 under platen cover (F).
3. (B5; B6) Latch, then press down on platen cover. Tighten screws (G). Remove 600 T 91030.

3. Latch, then press down on platen cover. Tighten screws (G) and spacers. Remove 600 T 91197. Put back interlock switch S3. Adjust, 16.2A Top cover interlock switch S3.
4. Loosen screw (A). Bias S15 towards the document glass. Ensure actuator arm on S15 clears roller on latch arm (B) by 0.13 mm (0.005 in). Tighten screw (A).
5. Loosen screws (C).

WARNING: IF SWITCHES S15 AND S2 DO NOT OPERATE CORRECTLY THE OPERATOR WILL BE EXPOSED TO THE DANGER OF 'FLASH'.

1. On B5 machines, loosen front and rear screws (G).
2. On B6 machines and later models, loosen front screws (G). Take off switch S3 (A1, Figure 16 Part 1) and loosen the spacers and screw that hold the rear hinge.
3. Put 600 T 91030 (B5) or 600 T 91197 (B6, B7) under platen cover (F).
4. Close and push down on the platen cover (F).
5. On B5 machines, tighten front and rear screws (G). Remove 600 T 91030.
6. On B6/7 machines, tighten screws (G) and spacers. Remove 600 T 91197. Install and adjust switch S3.
7. Loosen screw (A). Move S15 towards the platen glass.
8. Adjust so that there is a 0.13 mm (0.005 in) clearance between roller K, (16.1.B) on catch arm (B) and the actuator arm on S15. Tighten screw A, (16.1.A).
9. Loosen screws (C).

Figure 26.1

response that found the MCE version far easier to understand and use.

It is important to stress that the judgment in terms of final acceptability is not that of the originator or translator, but that of the end user, in our case the Service Representative. The judgment is based upon the end user's ability to follow the instructions easily and quickly with no negative impact on job performance.

With the computer programmed and the necessary vocabularies or dictionaries and target language loaded, the system is ready to go.

The source language text is fed into the computer and the target output can be delivered in either hard copy or displayed on a video display unit. The next stage is to post-edit the output. This involves identifying errors, analyzing them and determining the cause and, if possible, determine solutions to eliminate similar errors in future. These solutions might fall in one of several areas. It may be necessary to add to, or modify, the existing dictionaries. It might be necessary to alter the writing rules for future use or it could be that the computer software needs adjustment. Each of these actions has a cumulative effect, gradually taking the total process nearer to the minimum acceptable productivity targets set against the system. It can be seen that in the early stages of implementation of such a system, it is very much a question of "running in" the system.

Status

At this point in time, we are extremely optimistic that computer aided translation, using our input controls and based on SYSTRAN, can be used to significantly improve our translation function. We have already shown translation productivity gains of better than 4:1. This level of productivity includes the post-edit function related to computer translation. Evidence to date suggests we will improve this level of productivity as we continue to use the system and reap the cumulative effect of software and file improvement. So far our tests have been limited and "off line." The program that we are working on now is designed to test the total process. This process will involve on line author originating the source text using the writing rules and the MCE vocabulary. The text will then be run through the computer and post-edited by a qualified translator for the target language. Translation is only part of the total process of developing and implementing a Service Documentation system. As in other systems, there is little to be gained in speeding up part of the process if you leave a bottleneck in another part of the system. For example, there is little

gain in spending millions of pounds to build a motorway if all it does is speed up the traffic to the motorway exits and create a traffic jam at the intersection. So with our approach to translation, it is an integrated part of a total system.

Once the post-edit stage has been completed, the system will permit further productivity benefits. As all the text, both source and target language, is held in the computer we can electronically file it, update and modify it and print it out. By hooking our translation systems directly into a computerized text editing system, we can automatically select type face, size, etc. and compose the final page on a video screen. This greatly speeds up the total process and eliminates the relatively slow and expensive text creation and composition stages each language has traditionally required.

To date, we have carried out tests with English to French and Spanish translation. Other language pairs will follow, but in each case it is essential to ensure the end user of the target language is involved in the development process. One obvious example where it is essential to gain acceptance from the end user is where you are selecting one base language that is to be used in more than one country, e.g.

French—France, Canada, Switzerland, Belgium

Spanish—Spain, Latin America

Dutch—Holland, Belgium

Portuguese—Portugal, Latin America

In each of the above situations, the variations of both languages in each country are significant. However, it is possible to gain acceptance for a common vocabulary between countries by careful selection and discussion.

Exactly how we will finally install the system in terms of function and location is still under development considerations. The diagram (figure 26.2) shows the principal activities that we are "hooking" into the

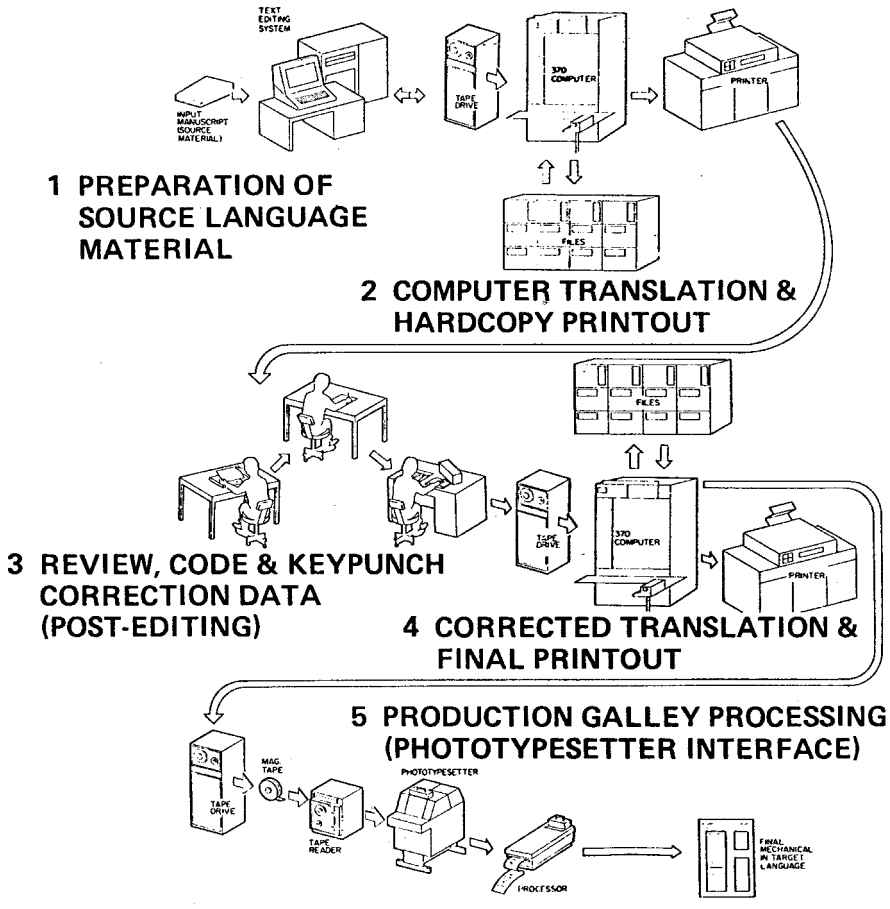


Figure 26.2

system. At present, the post-edit function is done from hardcopy, as it is an integral part of the input and computer software preparation. Once the system is up and running and post-edit becomes purely a translation/editing function, the work could be done remotely or on site, direct on a video display unit.

At this stage, we cannot say for what ranges of application computer aided translation will be suitable. As was said earlier, the system can be used for pure technical communication, where facts are listed for future retrieval. Whether it is possible to extend this into the area of training has yet to be established. Within the definition of training in our company, we range from pure technical skill-based training to the highly interpretive interpersonal skill training. It seems reasonable to assume that the more straightforward technical training offers the best opportunity to use computer aided translation. However, it must be appreciated that training materials are not written to record data but to enable initial learning. To this end, training material tends to be written in a more personalized style, making use of colloquial expressions and localized examples. Obviously, this type of translation requires a combination of the translator's skills and also those of the Trainer. In short, we are now in the field of interpretation, rather than translation and at this moment in time the computer falls short of that particular target. Already computer aided translation has come a long way, there is every reason to believe it will go still further. Our judgments on its acceptability must be based on realistic performance criteria and not on subjective argument. We are not trying to perfect an automated system that "appreciates" the finer points of a particular language, but a "tool" to assist us in the functional translation of a specific area of business communications.

This page intentionally left blank

III
SYSTEM DESIGN

This page intentionally left blank

INTRODUCTION

Harold Somers

In this section we concentrate on system design. There is of course something of an overlap between this and the previous two sections, since system design has from the very beginning been a methodological issue. However, in this section we present papers which have had an impact on the field especially from this viewpoint. One way that we might have represented this aspect of MT would be to present descriptions of particularly influential individual systems; but we quickly realized that, with a few notable exceptions, the choice of which systems to present would have been very difficult. Many systems in any case undergo a considerable change in design during the lifetime of their development, which might have caused us some difficulty. Indeed, some of the papers included here present system design issues from the viewpoint of a particular system or project, and the reader should bear in mind that they do not necessarily represent the eventual configuration for that particular project, as we shall see.

The first paper in the section is one of the earliest explicit descriptions of MT design. Predating the ALPAC report by six years, it belies the often-stated view that linguistic and computational sophistication came to MT only after that damning report. Michael Zarechnak's presentation to the June 1958 meeting of the ACM describes the Georgetown system which was a forerunner of Systran, perhaps the single most successful MT system, and presents an approach which was to become entirely familiar over the next 30 years. In his general analysis technique (GAT), implemented on an IBM 701 machine, we see one of the first examples of the separation of algorithms from the linguistic knowledge that they utilize, and in the three-level approach to linguistic analysis, we see the stratificational approach that became so widespread. Zarechnak gives details of both his linguistic method, which he calls morphemic, syntagmatic and syntactic analysis, and of the data-structures used by the program: necessarily crude but actually not all that different from structures still in use some 25 years later (e.g., SUSY—Maas 1987).

As we know, there was not much activity in MT in the late 1960s, and we jump almost 20 years, to COLING 1976 for our next landmark paper, Bernard Vauquois' survey of different approaches. This is the article in which the distinction between first-generation and second-generation architectures is made, and is possibly the first appearance of the famous "pyramid" diagram that is almost obligatory in any general article about MT. Notice by the way that the diagram originally appeared with the apex at the bottom, facilitating the metaphors of surface and deep representations, which seem somehow less intuitive when the diagram is inverted. The key elements of the second generation are all laid out here: the modularity and stratification of the translation process into analysis (parsing), transfer, and generation; the possible extrapolation of the analysis phase to an extent that transfer is unnecessary (the pivot or interlingua approach); the use of formalized and computable language models, and their nature (finite-state or context-free); and the separation of algorithms from the linguistic data. The paper ends with the suggestion that a third generation of MT systems would make use of the results of research in artificial intelligence, incorporating richer semantics, and some knowledge of the real world.

Vauquois also describes ways in which the parallel goal of human-assisted MT could be achieved. Although not often cited, this paper clearly set the agenda and defined the vocabulary of MT system design for the following 20 years.

The AI approach mentioned by Vauquois is well illustrated by Yorick Wilks' description of his MT system, developed at Stanford University in the early 1970s. Wilks' system was designed both to translate and understand text, the latter to be demonstrated by an ability to answer questions (though Wilks was later to claim that translation was often as good a test of understanding as any, especially if it involved resolving ambiguities of word-sense, syntactic structure, pronouns and so on). Wilks distances himself somewhat, in the opening paragraphs, from the formal logic approach to semantics that was prevalent at that time, and he places his approach firmly in the interlingua camp. Wilks describes his approach in a lot of detail, which was somewhat unusual at the time, and the paper is above all interesting in that Wilks illustrates and discusses explicitly his proposed interlingual representations, and his examples tackle a variety of ambiguities and other difficulties.

While Wilks and others explored the possibilities of MT systems influenced by advances in AI, Alan Melby took on the proposal to develop systems where the computer would cooperate with a human user to produce high-quality translations. Although this approach had been suggested by various commentators, notably of course the ALPAC report, but also Lippmann (1971) and Kay (1980), it was Melby who can be credited with having done the most to see these ideas realized. In a series of articles developing the theme, and in software which was eventually marketed commercially (in the form of the ALPS system), Melby's "Interactive Translation System" (ITS) became a blueprint for the Translator's Workstations that are now more or less familiar. Melby's key idea was that the computer should be flexible enough to offer aid to the translator at different levels ranging from simple text processing to terminology aids to full machine translation. Melby's thoughtful analysis of the role of the translator in MT and his personal experience of this job have had an important impact on the field.

The European Commission's EUROTRA MT project was, and perhaps always will be, the largest MT project ever undertaken, both in terms of cost and personnel. It is not controversial to say that its outcome was a huge disappointment, and this is not the place to discuss that aspect of it. In its early days, the project was shrouded in a veil of secrecy, imposed by the funders, so that few details of its design were published, beyond fairly banal and superficial descriptions of the impact on the system design of the organizational structure of the project (in particular, the geographical dispersal of those working on the project, and the desire to accommodate diverse scientific predilections). Reproduced here is an extract from the article which appeared in the 1985 special issue of *Computational Linguistics*, containing descriptions of more or less all the important MT systems at that time. The article was mostly about the general design and organizational structure, but the section reproduced here also shows that the project resulted in some innovative ideas about some computational aspects of MT system design. The extract discusses the problems of finding the appropriate level of specificity and generality for a linguistic formalism and implementing it in a distributed and robust fashion. The discussion illustrates the underlying tensions between procedural and declarative programming styles, providing a framework that was comfortable for linguists with varying experience of computational linguistics, the result needing also to be efficient and reliable. Although Johnson, King and des Tombe rejected the use of an existing programming language, perhaps extended by a library of purpose-built macros, subroutines or functions, eventually this was the approach adopted for the EUROTRA system, though it should be said that in the choice of Prolog for this task, many of the concerns and ideas expressed in this early article were influential.

Makoto Nagao has been one of the most influential and important names in MT research, not only in Japan but worldwide. The paper he delivered at a minor symposium in France in 1981, published three years later in a little-read collection, languished in obscurity until the start of the next decade, when suddenly and unexpectedly a whole new paradigm for MT emerged. Nagao's paper is inevitably cited as the first one in which Example-based MT is proposed, although actually Nagao does not use this term, but rather talks of "machine translation by example-guided inference," or "machine translation by the analogy principle." The main features of EBMT are there nevertheless: the use of examples rather than rules to establish the correspondences; and the need for some means to quantify the similarity between the input and the various examples (Nagao assumes the use of a thesaurus).

Apparently quite independently of Nagao, the BSO research group in Utrecht, and in particular Victor Sadler, had a number of ideas about using a small corpus of examples as a general-purpose knowledge source for NLP purposes. In including this paper in our collection, we are perhaps departing slightly from our goal of including influential and much-cited papers, since this one, presented at a semi-private (invitation-only) seminar, is probably not widely known. But we include it because it contains several ideas which were later to become widespread, and thus Sadler should be acknowledged as one of the first researchers to suggest them. For example, since the sentences in the corpus were stored as grammatically annotated tree structures, this is an early example of a tree bank. Sadler goes into extensive detail about how such a resource can be developed and used, using the term "example-based" explicitly, and probably predating the use by various Japanese researchers of that term. Interestingly, the seminar where this paper was presented was organized by ATR, one of the groups which is closely associated with this approach. The BSO group had already presented their idea of a bilingual knowledge bank, another analogical technique especially useful for word-sense disambiguation, at COLING in 1990. In fact, the BSO group never really got the opportunity to explore their ideas about EBMT fully, being victims of changed funding priorities in the mid-1990s.

Another new technique which emerged at the beginning of the 1990s was the "statistical" approach, with the IBM group led by Peter Brown in the forefront. The paper reproduced here appeared in *Computational Linguistics* and gives the most complete description of their early experiments, which had been presented at various conferences in the two preceding years, the first presentation to an MT audience having been at the TMI conference at Carnegie Mellon University, Pittsburgh, in 1988. In this article are the essential elements of the approach: a later article (Brown et al. 1993) gives more details about the mathematical models, and indeed the statistical approach itself was later modified to take more account of linguistic generalizations, e.g., morphology, before the group split up some six or seven years later. At the time, the statistical approach, along with EBMT, was seen (by some) as a serious challenge to the by now traditional rule-based approach, this challenge typified by the (partly engineered) confrontational atmosphere at TMI-92 in Montreal. Although some researchers are still following a strictly empiricist approach, the more significant outcome is now a number of hybrid system designs involving statistical, corpus-based and rule-based processes (see Somers 1999 for a review). The related activity, not strictly MT but somewhat relevant, of bilingual corpus alignment (e.g., Gale and Church 1993, Kay and Röscheisen 1993, Melamed 1996, Fung and McKeown 1997) has enjoyed a great deal of attention in recent years, and has contributed to the development of a number of useful tools for translators (e.g., Dagan and Church 1997, Macklovitch and Hannan 1998, Simard and Plamondon 1998).

Another trend for the 1990s is typified by our last two papers: dialogue translation. The first paper describes the design of a system which would translate on-screen dialogues between two partners. The paper mentions difficulties of distinguishing

user–user dialogue, to be translated, and user–system dialogue, since the system includes a module to negotiate with the user about the content of the dialogue. Somers et al. actually introduce a second theme, however, which also has proven to be a predominant one in the 1990s, and which they dubbed "translation without a source text." Adapting MT for users other than translators, and who may even be monolingual, multilingual generation of target text on the basis of negotiation with the user is presented. Subsequent system designs proposed variants which could be described as "multilingual summarization," where the source data, which may or may not be in a textual form, is analyzed and represented to the user in a variety of textual forms which are not necessarily based on that of the original.

The final paper represents what is perhaps the new frontier for MT system design: speech translation. For a long time, it was assumed that the difficulties of speech recognition and understanding combined with those of translation would ensure that speech translation remained a dream well into the next century. Even during the preparation of this collection, reported research on speech translation did not reach significant proportions until very recently. We therefore choose as a representative of this newest of approaches a paper describing the ATR project, which arguably set the pace for speech translation research by daring to attempt it. Admittedly, by restricting the domain and concentrating on a relatively small training corpus, the ATR group have made their task as "easy" as possible, but we can probably expect, in any second volume of *Readings in Machine Translation* that might appear, a significant number of papers tackling various aspects of this problem.

References

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. "The Mathematics of Statistical Machine Translation: Parameter Estimation." *Computational Linguistics*, 19, 263–312.

Dagan, Ido, and Ken Church. 1997. "*Termight:* Coordinating Humans and Machines in Bilingual Terminology Acquisition." *Machine Translation*, 12, 89–107.

Fung, Pascale, and Kathleen McKeown. 1997. "A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora Across Language Groups." *Machine Translation*, 12, 53–87.

Gale, William A., and Kenneth W. Church. 1993. "A Program for Aligning Sentences in Bilingual Corpora." *Computational Linguistics*, 19, 75–102.

Kay, Martin. 1980. "The Proper Place of Men and Machines in Language Translation." Xerox PARC Working Paper, Palo Alto, CA. Reprinted in *Machine Translation*, 12 (1997), 3–23, and in this volume.

Kay, Martin, and Martin Röscheisen. 1993. "Text-Translation Alignment." *Computational Linguistics*, 19, 121–142.

Lippmann, E. O. 1971. "An Approach to Computer-Aided Translation." *IEEE Transactions on Engineering Writing and Speech*, 14, 10–33.

Maas, Heinz-Dieter. 1987. "The MT System SUSY." In Margaret King (ed.), *Machine Translation Today: The State of the Art*. Edinburgh: Edinburgh University Press, 209–246.

Macklovitch, Elliott, and Marie-Louise Hannan. 1998. "Line 'Em Up: Advances in Alignment Technology and Their Impact on Translation Support Tools." *Machine Translation*, 13, 41–57.

Melamed, I. Dan. 1996. "A Geometric Approach to Mapping Bitext Correspondence." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, PA, 1–12.

Simard, Michel, and Pierre Plamondon. 1998. "Bilingual Sentence Alignment: Balancing Robustness and Accuracy." *Machine Translation*, 13, 59–80.

Somers, Harold L. 1999. "Example-based Machine Translation." *Machine Translation*, 14, 113–158.

sible context of a given Russian word, the computer is provided with a series of operations permitting exhaustive analysis of the unique context, and the resulting generation of diacritics indicating the behavior of a word within this unique context, the sentence being translated at the moment. We include in the mechanical glossary only the inherent characteristics of the Russian word. For example, if the word is a noun, its features will be coded in terms of its gender, palatalization, paradigmatic set, idiom participation, and semantic features. We list in the glossary only the base or stem of the noun, and thus avoid the redundancy involved in listing the noun in all its inflected forms.

Now let us turn to the details of the matrix format, in figure 27.1.

Section A contains the input data, taken from the Russian glossary, and located in rows 1 through 12.

Section B represents analysis level 1, the operation effecting morphemic analysis (putting grammatical suffix and stem together). The results of this operation are recorded in rows 13 through 17.

Section C is the second level of linguistic analysis. It contains the locations for storing codes pertaining to relations between immediately adjacent words on the basis of the discovered linguistic structures of agreement, government and apposition. All of these codes are generated by the computer program and stored in rows 18–26.

Section D is analysis level 3, the syntactic operation. When the subject of the sentence is located, an appropriate code is stored at this location. Furthermore, the cuts between noun phrase and verb phrase are registered here. The results of this operation are stored in rows 27–32.

Section E is the output working area, where the English equivalent is synthesized. The English stem is selected to replace the Russian word stem, and the Russian grammatical ending is replaced by an appropriate English ending or by the insertion of a preposition. The result is stored in row 33.

Any Russian word is subject to analysis at all three levels, but positive results will be recorded at only a portion of the vertical locations, depending on the nature of the given word.

We will now give concrete examples taken from sections B, C, D, and E.

The first is from section B, morphology.

Let us assume, for example, that the Russian input contains six letters. The first five of these are found in the glossary as a possible word stem. All six letters (the full form of the word) are not located in the stored glossary. The sixth letter is -E, which is found in the list of possible endings, or suffixes. At this point the ending E operation goes into effect and follows the sequence outlined in the flow chart in figure 27.2.

By way of explanation of the symbols used in the flow chart, the linguistic "parts of speech" are designated as follows: U-1: noun; U-2: verb; U-3: adjectival; U-4: adverbial; U-5: preposition; U-6: conjunction; U-7: particle; U-8: punctuation; U-9: non-Cyrillic forms (such as numerals, Romanized expressions).

The example to be presented from section C, the syntagmatic phase of the translation program, is a portion of the agreement operation. Agreement is one of the three linguistic structures which characterize immediately adjacent words. Let us describe briefly the nature of these three structures.

By government structure, we understand a state of predictability of the inflectional case of a second word, on the basis of the preceding case determiner in the first word. Take for example the choice between the forms *they* and *them* in the English sentence *I saw* _____ *this morning*. The native speaker will of course select the word *them*. Its form is said to be governed by the preceding verb.

By apposition structure, we refer mostly to the relationship of an adverbial form to some particular word in the sentence. The two together comprise a meaningful set, yet there is no formal grammatical relationship of government or agreement to mark the bond. An adverbial item in Russian can relate to a noun, verb, adjective, or another adverb. A similar situation exists within the English language; for example, in the sentence *I saw them briefly this morning*, where the *-ly* form as an adverb modifies the verb.

Now let us discuss agreement structure, from which a concrete example of the mechanical operation will be given. By agreement structure, we mean an identical distribution of some grammatical feature between two words. Compare the phrases *this young tree* versus *these young trees*. The words *this* and *these* are not mutually replaceable, nor are the words *tree* and *trees*, whereas the word *young* does not participate in this type of grammatical game. The common feature exhibited between *this* and *tree* and between *these* and *trees* is the concept of singular versus plural. In Russian the word *young* would also share this feature.

Many other combinations in addition to that of adjective and noun enter into agreement relationships, but we shall consider for the purposes of this

27
Three Levels of Linguistic Analysis in Machine Translation[1]

Michael Zarechnak

Since October 1956 a linguistic research project in Machine Translation has been in operation at the Institute of Languages and Linguistics of Georgetown University in Washington, D.C. Prior to the onset of this full-scale project, Georgetown University carried out with the International Business Machines Corporation the first practical computer test in machine translation; this experiment was conducted on an IBM 701 early in 1954. At the present time Georgetown is but one of several American universities and corporations sponsoring research in the field. Research is also being done in England and the U.S.S.R.

The Director of the Georgetown project, L. E. Dostert, has consistently encouraged diversity in approach to the problem of mechanical translation. It is believed by those who have worked in the area that there is no unique solution to machine translation. Within the Georgetown project, there are currently three different groups working on Russian-to-English machine translation, and work is also being done in French-to-English machine translation. One of the Russian-to-English groups has developed a general analysis technique based on the concept of structural transfer from the source to the target language. This approach is designed to effect a complete analysis of the linguistic structure and semantic content of the Russian input text; the use of this type of analysis is not limited strictly to English translation, but has application to such uses as information retrieval and translation into other languages. It is of this General Analysis Technique (nicknamed GAT) that I will speak here.

Although any method of translation, whether human or mechanical, requires the substitution of the words of one language for those of the other, the nature of linguistic structure precludes strict linear substitution. English words cannot be directly substituted for Russian words because the grammatical inter-relationships within the two languages are not identical. Problems of lexical (vocabulary) choice between multiple equivalents, of word or phrase rearrangement, of insertion and deletion, are some of the problems encountered when translating from Russian to English. The General Analysis Technique holds it necessary to view the translation operation in terms of a machine-programmable analysis and transfer of successively included constituents within the sentence.

The linguistic analysis can be characterized in three successive levels, or stages, which are effected internally by the computer between the input and output phases. What are these three levels? We will begin with a brief description of each, and then turn to concrete examples.

The first level concerns the analysis of the individual word. It may be inflected, meaning it may take variant grammatical endings. An example of this is given in figure 27.1.

The second level deals with relations existing between immediately adjacent words. The result of this analysis is a series of building blocks out of which the last level is constructed, namely the sentence. The types of building blocks for the sentence are contained within government, agreement and apposition structures.

The third level solves such problems as locating the nucleus of the noun phrase and verb phrase within the sentence. The first in most cases will be a noun in the nominative case or some substitution for it; the second takes the form of some type of verb or its substitution. This level secures enough information so that the English structural equivalent can be elicited.

In our linguistic jargon we refer to the first, second and third levels as morphemic, syntagmatic and syntactic, respectively.

These levels are not self-contained or independent stages; they represent segments of the whole machine translation technique as devised by my section of the research project. Inasmuch as language, just as any other phenomenon of the world we live in, exhibits regularity and patterning, I believe that the linguist can discover and describe the underlying concepts of this ordered system which we call language. The external expression of linguistic pattern is comparable to the time function; the irreversibility of the latter

Constant Locations		Content	Shifting Locations				
			Word				
Level	Row		1st	2nd	3rd	4th	Nth
A. Input	1	Russian word					
	2	Part of speech					
	3	Paradigmatic set					
	4	Gender					
	5	Idiom candidacy					
	6	English equivalent(s)					
	7	Transfer ambiguity					
	8	Case determiner					
	9	Animation					
	10	Time					
	11	Space					
	12	Voice					
B. Morphology	13	Number					
	14	Full form					
	15	Tense					
	16	Person					
	17	Case					
C. Syntagmatic	18	Interpolation					
	19	Class function					
	20	Homogeneous function					
	21	Apposition					
	22	Agreement					
	23	Noun }					
	24	Verb }					
	25	Prepositional } government					
	26	Adjectival }					
D. Syntax	27	Exclusion					
	28	Boundary					
	29	Independent variable					
	30	Dependent variable					
	31	Syntagmatic }					
	32	Syntactic } rearrangement					
E. Output	33	English word					

Figure 27.1
Matrix format.

is reflected in the importance of sequential analysis within the three levels. It is not surprising, then, that a linguist should develop the concept of a rectangular matrix to describe all the necessary operations in machine translation. (Of course I am aware of the pseudo-mathematical flavor of some of my statements, but from the linguistic point of view the matrix idea is a very practical device, exhaustive yet simple, and yielding the desired analysis.)

The rows of the matrix consist of constant operations, representing vertically for each word the operations necessary for the machine to produce all the codes to be used in translation.

The columns are shifting in character, in that the number of columns depends on the number of words in the sentence. In the Russian chemical corpus which we have used for analysis this number varies from 5 to 70 words.

The basic feature of the General Analysis method is the principle of computer-generated translation codes. Instead of the linguist supplying these in the Russian glossary, thereby having examined any pos-

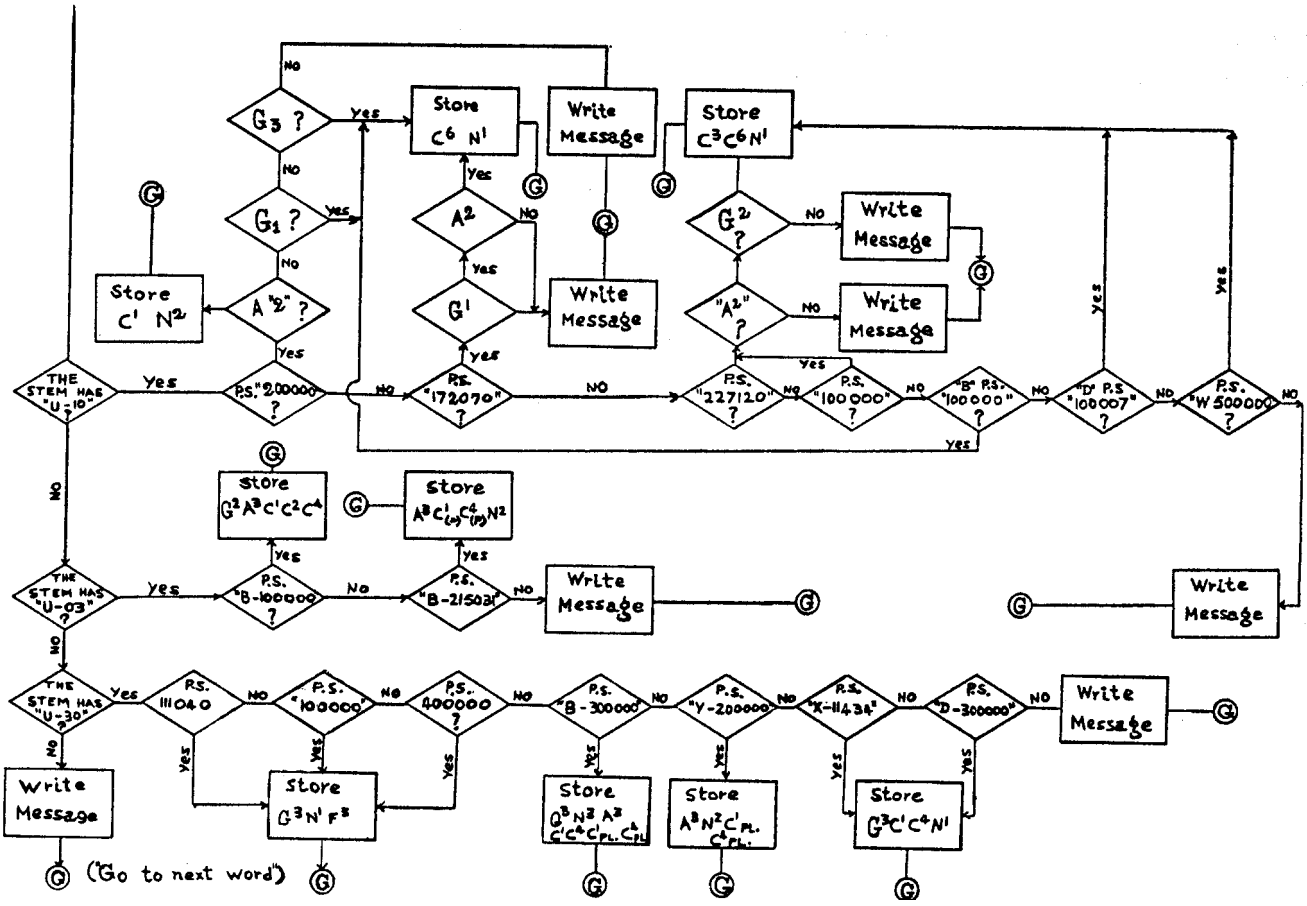


Figure 27.2
E operation.

discussion only the adjective plus noun set. Such pairs function as single units within the total sentence structure and need to be identified not only for purposes of grammatical translation but for operations of rearrangement for the English output.

The job of the computer program is to locate and identify agreement structures as they appear in the context of a Russian sentence. The program then attaches an appropriate diacritic to both members of the structure, and this diacritic is designed to indicate the nature of the agreement relation, the classes participating in the structure, and the grammatical features which control the relation.

The computer program proceeds as follows. It checks beginning with the first word of the sentence for the occurrence of an adjective. When a member of the adjective class is located, a check is made for a noun occurring immediately to the left or to the right. If so, the grammatical features of the adjective and noun are compared to discover whether they participate in an agreement relation. If all the necessary criteria are satisfied, a diacritic is stored at a particular address under each member of the structure. This is a four-digit diacritic. The first and second indicate the classes of the participating members. The third digit indicates the type of grammatical relationship, and the fourth records the inflectional case which characterizes the structure. For example, upon encountering the words *ximiceskix soedinenii*, meaning "chemical compounds", the computer will store under both words the diacritic 3112; 3, 1 means adjective plus noun, the third digit 1 means regular agreement, and the final digit 2 means the genitive case.

A flow chart for a portion of the agreement operation is given in figure 27.3.

The programmability of these linguistic formulations has been confirmed by several runs on the IBM 705 computer. Tests have included the idiom glossary look-up, and detailed syntagmatic and syntactic operations from levels C and D.

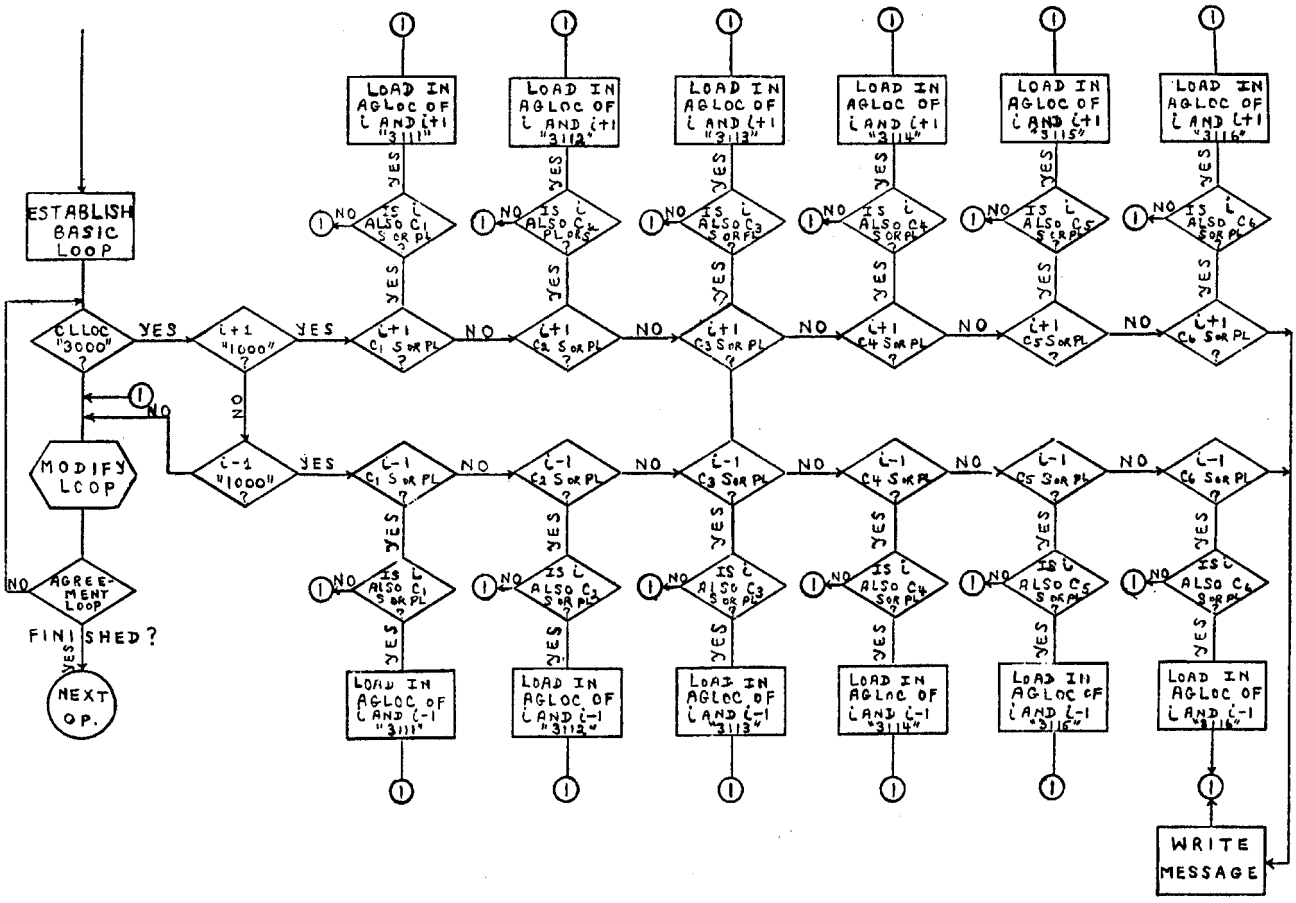


Figure 27.3
Agreement operation.

One sentence which has been analyzed in the partial tests is the following:

НА ЛИНИИ ЛИКВИДУСА СИСТЕМЫ, ИССЛЕДОВАННОЙ ДО 65 МОЛ. % КСЛ/ ДАЛЕЕ ИЗУЧЕНИЮ ПОМЕШАЛА ВЫСОКАЯ ТЕМПЕРАТУРА ПЛАВЛЕНИЯ СМЕСИ/, ИМЕЕТСЯ РЯД ВЕТВЕЙ КРИСТАЛЛИЗАЦИИ ИНКОНГ РУЭНТНО ПЛАВЯЩИХСЯ ХИМИЧЕСКИХ СОЕДИНЕНИЙ. (*J. General Chem., Moscow, 22* (1952)).

The code generated by the computer and stored under the words of the sentence were utilized by the program to produce the following English translation:

On the liquid curve of the system, studied up to 65 mol. % KCL (the high melting point of the mixture prevented further study), there is a series of branches of crystallization of incongruently melting chemical compounds.

The codes produced under each word are as in table 27.1.

Section D, the syntactic level, is designed for rearrangement operations within noun phrases and verb phrases as well as between the two. It is necessary for the computer to identify the head word of the noun phrase and the head word of the verb phrase. This routine makes possible a compression of any Russian sentence type into one of the following: 0-0, 0-1, 1-0, 1-1, 1-2, 2-1, 0-2, 2-0, or 2-2. The first digit of each set refers to the head word of the noun phrase, and the second digit to the head of the verb phrase. Zero means absence of the form, 1 means single occurrence, and 2 means "more than single occurrence." Thus a Russian sentence containing two subject noun phrases and one verb phrase is represented as the type 2-1. We refer to the head of the noun phrase as the independent variable, and to the head of the verb phrase as the dependent variable.

A flow chart for the operation which identifies the head word of the noun phrase is given in figure 27.4.

Table 27.1

NA	5000				5126			
LINII	1000			1122	5126			
LIKVIDUSA	1000			1122	5126			
SISTEMY	1000			1122	5126			
,								
ISSLEDOVANNOJ								
DO	5000				5122			
65	3000	3002	3112		5122			
MOL.	3000	3002	3112		5122			
%	3000	3002	3112		5122			
KCL	1000		3112		5122			
/								
DALEE	4000					413 P	E	
IZUCENIH	1000				2123	413 P	E	
POMEWALA	2000				2123		E	Pr
VYSOKA4	3000		3111				E	
D								
TEMPERATURA	1000		3111	1122			E	H
PLAVLENI4	1000			1122			E	
SMESI	1000			1122			E	
/								
,								
NA	5000				5126			
IMEETS4	2000			1122				Pr
R4D	1000			1122				H
D								
VETVEI	1000			1122				
KRISTALLIZAQII	1000			1122				
INKONGRU3N TNO	4000					433 P		
PLAV45IXS4						433 P		
XIMICESKIX	3000		3112					
SOEDINENII	1000		3112					

Finally, we present an example from the transfer procedure, to demonstrate how semantic criteria are used in this phase. We store with each word three semantic cues, if these are inherent in the word. Thus the preposition *DO* may be translated into English in different ways, depending on certain semantic criteria of time and space in the immediate context of the preposition.

Figure 27.5 is the flow diagram for the translation of the preposition *DO*, indicating the method of choice between multiple equivalents in English.

In conclusion, I would like to make a few remarks concerning current planning for continued test runs on the computer. We expect to translate a continuous corpus of more than 1000 sentences before the end of the calendar year. If this translation is successful, we can rapidly increase the scope of machine-translated Russian scientific material, since our dictionary look-up is not complicated and the addition of new words will not demand any change in the basic translation routine. A greatly expanded corpus may require the addition of some new operations covering certain structural features which have not occurred in the initial corpus. Because the formulation has been done on the basis of generalized linguistic concepts of Russian structure, we do not expect any radical changes in the existing program, no matter how many sentences we put to the test.

Note

1. Presented at the meeting of the Association [for Computing Machinery], June 11–13, 1958.

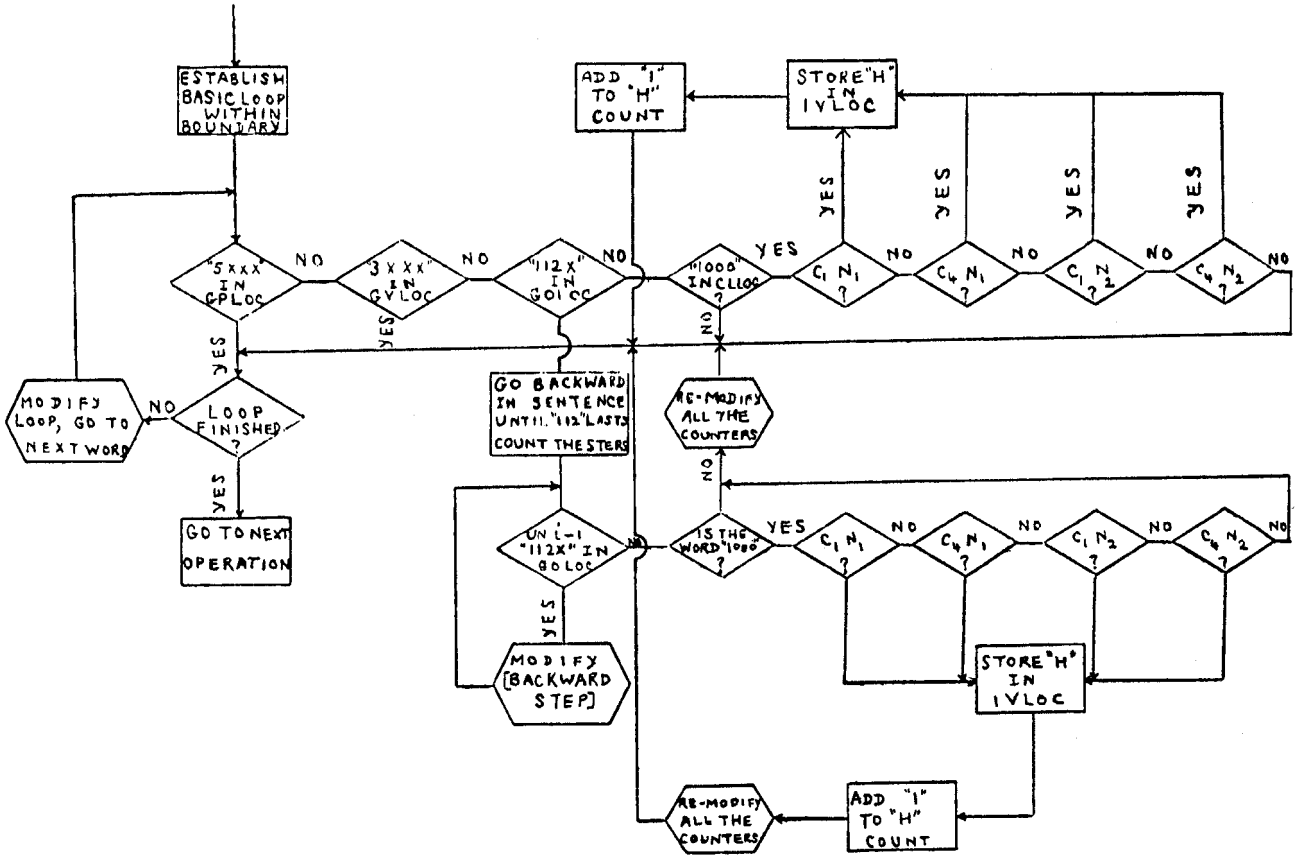


Figure 27.4
Subject operation.

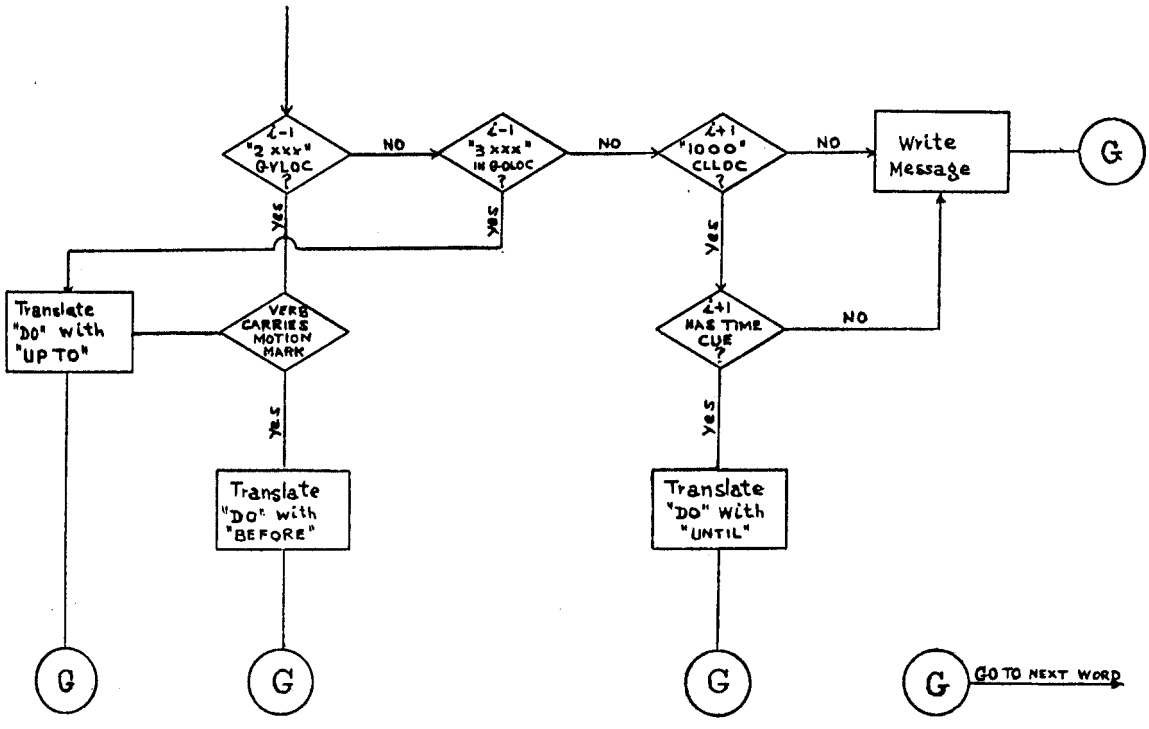


Figure 27.5
Do operation.

This page intentionally left blank

28
Automatic Translation—A Survey of Different Approaches

B. Vauquois

Origin and Motivations of Automatic Translation

As this international conference COLING 76, like the others since 1965, is devoted to "Computational Linguistics," it may be a good opportunity to recall the role of automatic translation in the development of this field.

In the beginning, when Y. Bar-Hillel at MIT was the first full-time researcher, the motivation for automatic translation was curiosity. The use of computers was almost entirely restricted to computation in numerical analysis; few scholars were thinking of other activities. Translating from one natural language into another with a computer appeared to be a feasible and very attractive task. At that time (1951–53), automatic translation was the most important subject (perhaps the only subject) in the field of what has since been called "computational linguistics." During the following years, many laboratories in different countries were created to survey and experiment in the new area.

Moreover, beside the original curiosity, an increasing demand for translation brought a practical goal to automatic translation which was believed to be a powerful and economical substitute for human translation. In fact, for almost fifteen years, this need for translation was considered exclusively for use in information gathering (for reading scientific and technical literature as well as newspapers published in foreign countries in their own language).

This is the main reason why in the United States and later on in Great Britain the target language was invariably English, in U.S.S.R. it was Russian, and in France and Japan it was French and Japanese respectively.

In the early sixties the situation was the following:

On the one hand, experimental designs for automatic translation systems had been checked on some pairs of languages with a limited corpus as one of these, more developed, was almost ready to provide the users of translation with a large program, using enormous dictionaries, for translation from Russian into English. This was the system developed by Georgetown University which can be considered as the leader of what we call the "first generation" of automatic translation systems. In fact, this system, which first became operational in 1963, is still in use at the Atomic Energy Commission, Oak Ridge (Tennessee), and at the Euratom Common Research Center, Ispra (Italy). Other similar programs have been derived from this initial work at Georgetown.

On the other hand, for many scholars, automatic translation was rather considered as a source of inspiration for more academic studies. Realizing an automatic translation system became a long-range project; and systematic research, both in linguistics (analysis, generation, and comparison of languages) and in computer science (formal models of languages, algorithms for parsing, adequate programming languages) took precedence over all considerations of utilization. That attitude reflected the research priorities followed by the so-called "second generation."

Characteristic Features of the First Generation

We have to keep in mind that the purpose of all programs characteristic of this generation was practical automatic translation, available as soon as possible. Furthermore, the period of designing such programs spreads from 1955 to 1960, at a time when linguistic models and their formalization were not very helpful and the software available offered poor facilities. Also, at the beginning of automatic processing of natural languages (devoted also to lexicography and quantitative linguistics), only the medium of punched cards was used. So, the activity of encoding extended to the level of processing by computers where programming became a subtle manipulation with positional codes for linguistic features which were chosen *a priori*. Consequently programs reflected the complete heuristics of the designer effectively including the grammar of the language by means of hierarchical questions represented by flow-charts.

The basic component of such a first-generation system is the dictionary which furnishes all lexical and

syntactic information, and the translation (or multiple translations) in the target language for each entry in the source language. Given an input text, the first step in such a system is dictionary look-up. In most cases this operation is performed by matching the form in the dictionary with the unanalyzed occurrence in the text. However, in the case of idiomatic expressions or particular strings of words, the longest match is usually preferred. For those source languages having many inflected forms for a single word, an elementary morphological analysis is sometimes performed by cutting the occurrence into a stem and an ending. After this dictionary look-up, each occurrence (or sequence of occurrences) is replaced by the information found in the corresponding entry of the dictionary. The next step consists of solving lexical ambiguities. Among the many kinds of ambiguities, the highest priority is given to the multiple syntactic class. For instance, for a word *matches*, it has to be decided whether it is a noun or a verb. A table of all specific syntactic ambiguities has been previously constructed, and for each case (verb–noun, verb–noun–adjective, verb–conjunction) an appropriate sequence of questions about the proceeding and following words, represented in the computer as a subroutine, has been established to find the correct solution. Some difficulties appear when many such ambiguities occur in the same sentence; the order of application of the different subroutines is sometimes important to avoid either the loss of alternate solutions or a blocking of the system. Nevertheless, the translating process goes on and consists of applying a sequence of translation routines dealing with words or groups of words. These routines reorder the words based on restricted context; many of them are called by specific lexical entries. Finally, if it can be solved within the selected context, a morphological routine computes the grammatical agreement (for instance, verbal conjugation) and morphological alterations.

In conclusion, the strategy of the first-generation system is based on a catalogue of linguistic facts which are locally relevant for a given pair of languages considered from the point of view of translation in one direction. The major guide for the composition of this catalogue and also for its use is the designer's knowledge of grammar and the experience of human translation. More sophisticated cases are solved, when possible, either by *ad hoc* subroutines (one for each case) or by direct translation in the dictionary, considering each such case as an extension of an idiomatic expression.

Characteristics of the Second Generation

As early as 1957, V. Yngve proposed "a framework for syntactic translation" [Yngve 1957]. The basic concepts of such a framework can be stated as follows:

First, the fragment of text considered as a whole is the sentence. Then, it is assumed that for each language a sentence may be adequately described by a structural specifier; in fact, only limited indications about the kind of information required for such a specifier were provided as it was too early to define an adequate formalization. Nevertheless, the idea arose of a system proceeding in these three general steps. This is the first characteristic of second-generation systems.

input sentence $\xrightarrow{1}$ structural specifier in the source language $\xrightarrow{2}$ structural specifier in the target language $\xrightarrow{3}$ output sentence

Step 1 deals only with the source language. It is the analysis (parsing) procedure; step 3 deals only with the target language (generation from the specifier); step 2 involves both languages at some abstract level (for the moment, transfer is restricted to syntactic structures).

The immediate consequences of such an approach were very important; in spite of a large amount of investigation and fruitful development for many years, the research still continues in this framework, and further results are expected. Indeed, this strategy matches the theory of stratification in natural languages and is well suited to the realization of computable models.

By a simple extrapolation, we can imagine as many levels as we wish from the zero level (the level at which the text considered as a string of characters), asymptotically towards a level of understanding. At each level a formalization of the input sentence can be defined. Then, it may be assumed that the deeper the level chosen, the easier the transfer is. At the limit, if the ideal level of understanding could be reached for a given sentence in one language, the same structural specifier would represent all the paraphrases of this sentence in all languages.

During the 1960s many laboratories worked within this framework, the selected level for transfer being more and more ambitious (from surface syntactic structure, to deep syntactic structure, sememic level, approximations of pivot languages, and so on) (see figure 28.1).

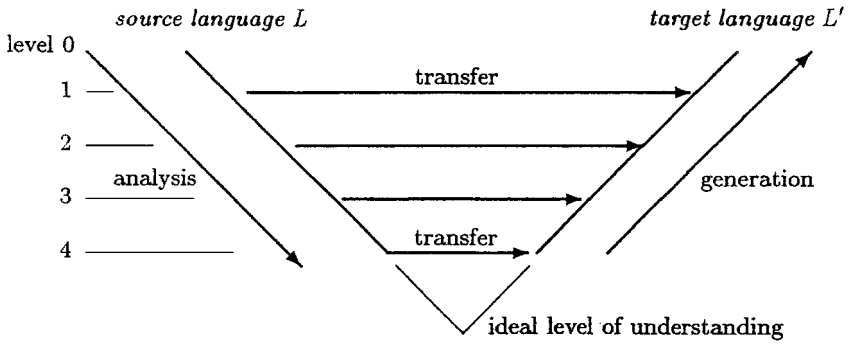


Figure 28.1

The second characteristic feature of this second generation concerns the way of representing linguistic data and the algorithmic approach. The stratification description of natural language implies a kind of representation at each level by means of some artificial language; then, access to level $n+1$ from level n in analysis (or level n from level $n+1$ in generation) needs a transduction from one artificial language to another. For such a purpose, the concept of "model," based on formalized and computable languages, appears to be fundamental. Also, the notion of artificial language, considered as a set of strings over some vocabulary, has to be extended to a set of rooted-trees as soon as deep levels are investigated.

By increasing the degree of complexity, we can say that from level 0 to morphemic level a finite state string-to-string transducer is powerful enough to ensure a perfect matching of the model to the linguistic reality; then a context-free parser has been used almost everywhere to simulate a string-to-tree transducer. The results of these parsings match for the first time the idea of structural specifiers assigned to sentences of the input text; even if a context-free model is not adequate for some sentences, the approximation is interesting.

Some laboratories performed automatic translations, at an experimental stage, using this level for transfer. This was the case in experiments conducted at the University of Kyoto and the University of Kyushu. The autonomic division of the National Physical Laboratory (Teddington) had a similar project. At a higher degree of sophistication, many attempts have been made to extend the power of context-free models slightly in order to reach a higher level of adequacy for the structural specifiers. Interesting results have been obtained in many places, including Rand, Harvard, IBM, Leningrad, Moscow, and Grenoble. Finally, by the end of the 1960s and in the early 1970s, automatic translation had been achieved by transfer at a deeper level. This is the case with the Grenoble system, which introduces "pivot languages," and the Montreal system, which identifies relations between words and syntagms at a sememic level. The latter group has recently built a system for translating weather forecasts from English into French for the Canadian meteorological network. This is the only example of a second-generation automatic translation system in use beyond an experimental demonstration.

A third characteristic feature of this generation concerns the way of programming. The different context-free parsers and other transducers are generalized routines for which grammars and dictionaries are considered as data, along with the text to be translated. Programs written for the first generation applied linguistic information directly to the input text. In contrast, the second-generation parsers and synthesizers compile for each grammar a program which operates in turn on the text.

The New Look of Automatic Translation

At the present time, a few automatic translation systems of the first generation are available; only one of the second generation is running (Montreal) and few others are expected in the near future (Leibniz group).

Research on a third generation began a few years ago. In contrast with the past, when we could seldom predict either what would be available or when, we now have a better idea of what can be expected in the near future (within 2–3 years) and in the more distant future (5–7 years).

For a long time, automatic translation has been torn between two opposite goals: a concrete and efficient system for commercial use on one hand, and on the other, scientific research on computational

linguistics. It seems now that some concrete applications are being obtained from the scientific side. Certainly, "fully automatic high quality translation" is not reachable in any foreseeable future, but we can expect feasible systems in which the process of translation is shared between human translator and computer. Before surveying the current trends which are developing in different places, we have to mention the new motivations for automatic translation research and production.

About twenty years ago, the primary demand was exclusively for translation for information gathering; this motivation still persists with an increasing intensity; but in addition, for the last few years, translations have been needed more and more for the dissemination of information. This is particularly true of multilingual organizations (for example, the European Community, and bilingual agencies in Canada). Certainly, the possibilities of human translation cannot face the totality of this demand any longer; the only solution lies in computerized systems, in which the knowledge amassed in the past fifteen years from the scientific approach should bring positive results.

If we consider what has been done within the framework of the second generation and what could bring immediate results in artificial intelligence, we can see how to satisfy simultaneously the requirements of scientific research and the expected demands for translation.

The artificial intelligence approach to natural language processing is mainly (if not exclusively) oriented toward the semantic interpretation of texts. This does not mean that all of the second-generation systems were restricted to syntactic description without any semantic considerations. On the contrary, the most sophisticated systems included some "semantic features" assigned to the lexical entries of their dictionaries, and these features were used by their grammars for "semantic agreement," in the same way as syntactic codes were used for grammatical agreement. In other words, the grammar rules were stated in terms of conditions on the combinatorial properties between classes which could be semantic classes as well as grammatical classes. More ambitious are the goals of semantic computation in artificial intelligence. Inference rules should be applicable to deduce new statements. The semantic consistency of a sentence or of a sequence of sentences should be open to evaluation. Furthermore, if a data base consisting of an appropriate description of knowledge about "the world" is stored in the computer, then any part of the input text (or deductions computed from it) should be open to evaluation for consistency with respect to that data base.

This approach seems to be the only means of solving the remaining ambiguities occurring at the end of syntactic analysis (even strengthened by semantic features); in particular, the reference of a pronoun outside the sentence cannot be determined by a sentence-to-sentence translation.

The introduction of such a semantic component into an automatic translation system is the characteristic feature of the so-called "third generation." However, all experiments conducted within this artificial intelligence approach are restricted to "microworlds;" it is too early to consider seriously a generalized use of this method, given the amount of information about the world needed for a translation of large amounts of text. But research into this third generation certainly needs to be extended.

Considering now the feasibility of automatic translation systems which merge human translators and the computer in a hybrid process, we can imagine several different strategies.

Let us assume a large translation service (about 1500 translators and revisors) where most of the texts have to be translated into several languages. In such a case, it seems, at least for the moment, that the work of the human translators must be separated from the work of the computer.

The complete process would be first pre-editing of the text (by inserting specific disambiguating markers); then, the automatic translation system operates on this edited text; and finally the output produced by the computer is revised to get the desired quality of translation. Of course, the balance between pre-editing and revision must be optimized according to the following considerations: ambiguities in the source text which are not solvable by the automatic translation system, the quality which is desired for the final result and the number of target languages. It is necessary to find a pre-editing code that is flexible enough to ensure such a balance and is compatible with the automatic device.

Another strategy would be machine translation aided by human translator in a conversational way. It is certainly the ideal way for the future. It would be interesting to develop experiments on a small scale (with a few users in time sharing) to improve the best ways of achieving this.

In both cases, as far as the automatic system is concerned, several remarks about the way of progressing can be enumerated:

1. The third-generation approach can be considered as an extension of the second-generation systems with a greater ambition concerning the transfer level (by contrast, first-generation systems cannot be assimilated). So, there is no obstacle to building a second-generation system with the possibility of increasing its power every time suitable progress in artificial intelligence is made.

2. The stratified models of the second generation are extremely rigid; the system applies these models in sequence, without any interference between them. A new approach consists of an arrangement of the grammars in such a way that the computation of the structural specifier at each level is not strictly sequential, but can be influenced by the others. A formalized representation of specifiers sharing the same graph structure but identified by different labels on the nodes was proposed by the automatic translation section of the Leibniz group two years ago.

3. As a consequence of the preceding remark, the usual parsers of the 1960s (responsible for this rigidity) are not suitable any longer. Among the new software systems designed for flexible computation at different levels, let us mention M. Kay's system [Kay 1973], Woods' extended transition networks [Woods 1970, 1973], Q-systems [Colmerauer 1971], REZO [Stewart 1978] and TARZAN [?] developed at Montreal, and ATEF, CETA, SYGMOR at Grenoble [Chauché 1975, Jaeger 1978].

In conclusion, the activity in automatic translation was the subject of renewed interest just a few years ago. All the progress realized in the processing of languages and in specialized software will contribute to successful realizations in the near future.

References

[Note: All the references in this paper have been added by the editors.]

Chauché, J. 1975. *Présentation du système CETA*. Rapport de recherche No. G-3100-A, GETA, Grenoble.

Colmerauer, A. 1971. *Les systèmes-Q: ou un formalisme pour analyser et sythétiser des phrases sur ordinateur*. Publication interne No. 43, TAUM, Université de Montréal.

Jaeger, D. 1978. *SYGMOR*. Rapport de recherche, GETA, Grenoble.

Kay, M. 1973. The MIND System. In R. Rustin (ed.), *Natural Language Processing*, New York: Algorithmics Press, 155–188.

Stewart, G. 1978. Spécialisation et compilation des ATN: REZO. Paper presented at COLING-78, Bergen.

Woods, W. 1970. Transition Network Grammars for Natural Language Analysis. *Communications of the ACM*, 13, 591–606.

Woods, W. 1973. An experimental parsing system for transition network grammars. In R. Rustin (ed.) *Natural Language Processing*, New York: Algorithmics Press, 111–154.

Yngve, V. H. 1957. A Framework for Syntactic Translation. *Mechanical Translation*, 4, no. 3, 59–65. Reproduced in this collection.

This page intentionally left blank

29
Multi-level Translation Aids

Alan K. Melby

Overview

At COLING-80, we reported on an Interactive Translation System called ITS (Melby et al. 1980). We will consider three problems in the first version of ITS: (1) human factors, (2) the all or nothing syndrome, and (3) traditional centralized processing.

The first problem (human factors) is the problem of keeping human translators and revisors happy. Humans naturally want to feel that they are doing useful, interesting work and that they are using the machine instead of it using them. However, the first version of ITS forced them to answer many uninteresting questions and to revise many sentences they thought should be retranslated.

The "all or nothing" syndrome is a name for the attitude that the machine must translate every sentence or it is not worth using a machine at all. The problem is that a system based on this approach is likely to be hard to adjust into a useful form if it does not attain the desired level of performance.

The problem with traditional centralized processing is that it does not provide consistent, reliable response time to each user and requires physical proximity or high-speed telecommunications. And a centralized system may be hard to decentralize after it has been designed.

The first version of ITS had all three of the above problems. These problems would disappear if we had FAHQT (Fully Automatic, High Quality Translation [Bar-Hillel 1960]). In that case a source text would be presented to the computer, which would promptly produce a polished translation, typeset and ready to be published without revision. That would solve the human problems because no human translators would be involved. The "all or nothing" question would be irrelevant because we would have it all. And centralized processing would not be a problem because there would be no interactive processing. This paper assumes that FAHQT of general text is not on the visible horizon and proposes a design which answers these problems.

In the new version of ITS, each translator works at a microcomputer instead of a conventional terminal. The microcomputers are part of a distributed network but can function without being on-line. The translator uses the microcomputer as a tool for getting the translation done and is in control of the translation process. There are three levels of aid available to the translator, ranging from simple text processing to terminology aids to full machine translation. All three levels are fully integrated and the translator can quickly switch from one level to another, even within the translation of a single sentence. This means that the translation process can continue smoothly regardless of how many sentences fail to receive a full analysis and a good machine translation. This in turn means that the actual machine translation component can be "pure" in the sense that no compromises need be made to ensure some kind of output even on sentences that are not analyzable with the current parser and model of language.

It is hoped that the above design will solve the three problems under discussion. Placing the translators in control of the operation of the system should improve their attitude. Using multiple levels of aid should overcome the dangers of the "all or nothing" approach. And replacing conventional terminals with microcomputers should overcome some of the problems of centralized processing. Solving these user-oriented problems is important from a theoretical viewpoint because even a research translation system desperately needs user feedback from real translators. And real translators will not give the needed feedback unless the system is practical and user-friendly.

The rest of the paper will elaborate on the three problems and their proposed solution in the new version of ITS.

Problem One: Human Factors

Lacking FAHQT, human translators and revisors are still needed in a computerized translation system. In ITS version one, translating a text involved asking

questions about each sentence of the text before the translation of the first sentence appeared. When the translated sentences finally did appear, the translator/revisor was expected to examine and then revise them as needed but not to retranslate them from the source text. After all, this was a human-assisted *machine* translation system and we had already invested considerable interaction time and machine time in the translation of each sentence. The translator/revisor was to remove the errors from the machine's translation and no more. Understandably, the human translator/revisor often felt more like a garbage collector than a translator.

Having an unhappy translator is a serious problem. It should be remedied, if possible, for two reasons: (1) We should be concerned for the translator as a person. (2) An unhappy translator will fight the system. Consider the following statement by a human translator:

> During my years with JPRS ... I had occasion to do some post-editing of machine translations, in addition to my normal assignments.... Monetary considerations aside, the work was odious. To post-edit, a conscientious translator had to literally retranslate every sentence in the original, compare it word for word with the clumsy machine attempts, and then laboriously print in corrections between the lines of the printout. It would have been much faster—and less tedious—just to translate "from scratch" and dictate the translation on tape, as I normally do. And I am sure the product would have been better. It was thus my impression that post-editing of machine translations is translation work at coolie wages. I can't imagine anyone wanting to do it unless the alternative was starvation. (Silverstein 1981)

Seppänen (1979) claims that relatively little attention has been paid to the pragmatic aspects of man/machine dialogues. He claims that human factors in man/machine interfaces have not attracted the interest of either computer scientists or psychologists. Perhaps, then, human factors in computerized translation systems are an appropriate area of interest for computational linguists, and this view seems to be gaining momentum from within the field. Researchers at the Grenoble project have concluded:

> The human and social aspects should not be neglected. To force a rigid system on revisors and translators is a guarantee of failure. It must be realized that [machine] translation can only be introduced step by step into some preexisting organizational structure. The translators and revisors of the EC did not only reject Systran because of its poor quality but also because they felt themselves becoming "slaves of the machine," and condemned to a repetitive and frustrating kind of work. (Boitet et al. 1980)

Our answer to the problem of human factors is to place the translator in control. The translator uses human judgment to decide when to post-edit and when to translate. Nothing is forced upon the translator. This approach is strongly argued for by Kay (1980) when he states: "The kind of translation device I am proposing will always be under the tight control of a human translator." And Lippman (1977) describes a successful terminology aids experiment in Mannheim and concludes: "The fact that quality was improved, rather than degraded as in the case of MT, appears to support the soundness of an approach where the translator retains full control of the translation process."

Problem Two: The "All or Nothing" Syndrome

Originally, FAHQT was the only goal of research in machine translation. Until recently, there seemed to be a widely shared assumption that the only excuse for the inclusion of a human translator in a machine translation system was as a temporary, unwanted appendage to be eliminated as soon as research progressed a little further. This "all or nothing" syndrome drove early machine translation researchers to aim for FAHQT or nothing at all. It is now quite respectable in computational linguistics to develop a computer system which is a *tool* used by a human expert to access information helpful in arriving at a diagnosis or other conclusion. Perhaps, then, it is time to entertain the possibility that it is also respectable to develop a machine translation system which includes sophisticated linguistic processing yet is designed to be used as a tool for the human translator.

If you expect each sentence of the final translation to be a straight machine translation or at worst a slight revision of a machine translated sentence, then you are setting yourself up for a fall. Remember Brinkmann's conclusion that

> the post-editing effort required to provide texts having a correctness rate of 75 or even 80 percent with the corrections necessary to reach an acceptable standard of quality is unjustifiable as far as expenditure of money and manpower is concerned. (Brinkmann 1980)

Thus, a strict post-edit approach must be nearly perfect or it is almost useless. Many projects start out with high goals, assuming that post-editing can surely rescue them if their original goals are not achieved. Even post-editing may not make the system viable.

The proposed solution to this problem is to anticipate from the beginning that not every sentence of

every text will be translated by computer and find its way to the target text with little or no revision. Then an effort can be made from the beginning to provide for a smooth integration of human and machine translations. ITS version two will have three integrated levels of aid under the control of the translator. We will now describe the three levels of translator aids.

Level one translator aids can be used immediately even without the source text being in machine-readable form. In other words, the translator can sit down with a source text on paper and begin translating much as if at a typewriter. Level one includes a text processor with integrated terminology aids. For familiar terms that recur there is a monolingual expansion code table which allows the user to insert user-defined abbreviations in the text and let the machine expand them. This feature is akin to the macro capability on some word processors. The key can be several characters long instead of a single control character, so the number of expansion codes available is limited principally by the desire of the translator. Level one also provides access to a bilingual terminology data bank. There is a term file in the microcomputer itself under the control of the individual translator. The translator also has access to a larger, shared term bank (through telecommunications or local network). Level one is similar to a translator aid being developed by Leland Wright, chairman of the Terminology Committee of the American Translators Association. Ideally, the translator would also have access to a data base of texts (both original and translated) which may be useful as research tools.

Level two translator aids require the source text to be in machine-readable form. Included in level two are utilities to process the source text according to the desires of the translator. For example, the translator may run across an unusual term and request a list of all occurrences of that term in that text. Level two also includes a suggestion box option (Melby 1981a,b) which the translator can invoke. This feature causes each word of the current text segment to be automatically looked up in the term file and displays any matches in a field of the screen called the suggestion box. If the translator opts to use the suggested translation of a term, a keystroke or two will insert it into the text at the point specified by the translator. If the translator desires, a morphological routine can be activated to inflect the term according to evidence available in the source and target segments.

Level three translator aids integrate the translator work station with a full-blown MT system. The MT component can be any machine translation system that includes a self-evaluation metric. The system uses that metric to assign to each of the translated sentences a quality rating (e.g., "A" means probable human quality, "B" means some uncertainty about parsing or semantic choices made, "C" means probable flaw, and "D" is severely deficient). On any segment, the translator may request to see the machine translation of that segment. If it looks good, the translator can pull it down into the work area, revise it as needed, and thus incorporate it into the translation being produced by the translator. Or the translator may request to see only those sentences that have a rating above a specified threshold (e.g., above "C"). Of course, the translator is *never* obliged to use the machine translation unless the translator feels it is more efficient to use it than to translate manually. No pressure is needed other than the pressure to produce rapid, high-quality translations. If using the machine translations makes the translation process go faster and better, than the translator will naturally use them.

The successful METEO system by TAUM (Montreal) expresses the essence of this approach. All sentences go into the MT system. The system evaluates its own output and accepts about 80 percent of the sentences. Those sentences are used without post-editing. The other 20 percent are translated by a human and integrated into the machine-translated sentences. This application differs from ours in that human translators do not see any machine translations at all—good or bad. But the basic level three approach is there.

One positive aspect of this three-level approach is that while level three is dramatically more complex linguistically and computationally than level two, level three appears to the translator to be very similar to level two. Level two presents key terms in the sentence; level three presents whole sentences. When good level three segments are available, it can speed up the translation considerably but their absence does not stop the translation process. Thus, a multi-level system can be put into production much sooner than a conventional post-edit system. And the sooner a system is put into production, the sooner useful feedback is obtained from the users.

The multi-level approach is designed to please (a) the sponsors (because the system is useful early in the project and becomes more useful with time), (b) the users (because they are in control and choose the level of aid), and (c) the linguists and programmers (because they are not pressured to make compromises just to get automatic translation on every sentence).

Problem Three: Traditional Centralized Processing

Machine translation began in the 1950s when the cost of a CPU prohibited the thought of distributed processing in which each user has a personal CPU. Interactive time-shared computing (where each user has a dumb terminal connected to a shared CPU) can give the impression that each user has a personal computer—so long as the system is not loaded down. Unfortunately, systems tend to get loaded down. Highly interactive work such as word processing is not suited to an environment where keystroke response times vary. Also, centralized processing requires either physical proximity to the main CPU or telecommunications lines. High-speed telecommunications can be very costly, and low-speed telecommunications are not user-friendly. A costly solution is to obtain a dedicated mainframe and never load it down. A more cost-effective solution in terms of today's computer systems is a distributed system in which each translator has a microcomputer tied into a loose network to share resources such as large dictionaries.

The individual translator work station would be a microcomputer with approximately 256 K of main memory, dual diskette drives, CRT, keyboard, small printer, and communications port. Such systems are available at relatively low cost (under $5000 U.S.). Additional storage for term files and text files can be obtained at reasonable cost by adding a Winchester-type disk. If several translators are in the same building, a local network can be set up to share terminology and document data bases and even inter-translator messages. The capabilities of the work station would include rapid, responsive word processing and access to internal dictionaries and to shared translator data bases (i.e. level one and level two processing). The internal dictionaries would include an expansion file and a terminology file under the control of the translator. Of course, the translator could load internal files appropriate to the subject matter of the document by inserting the appropriate diskettes. Access to source texts, document-specific dictionaries, and level-three machine translations could be granted through a local network, a telecommunications network, or through the mail on diskette. Ideally, part of the machine translation would be done on the translator work station in order to allow the translator to repair level-three dictionary problems before they cause repeated errors throughout the text. A minimal capability in the work station would be a translator-defined replacement table to correct some improper word choices that cause repeated errors in the machine-translated sentence. Ultimately, microcomputers will be powerful enough to allow source text to be presented to a work station which contains full level-three software. In the meantime, the raw machine translation part of level three can be done remotely on any suitable mainframe and then transmitted to a microcomputer translator work station for integration into the translation process as level-three aids.

Conclusion

The system described is not, of course, entirely original. It draws on ideas from university colleagues and others such as Kay, Boitet, Lippman, Andreyewski, Wright, and Brinkmann. But it does represent an important shift in direction from past years of research on ITS at Brigham Young University.[1] It is an integration of a machine translation system and a terminology aid system, with the final translated text being produced on a microcomputer in a distributed network.

The author's major motivations for pursuing this system are to provide a useful translator aids system and to create an appropriate vehicle for machine translation research. Fortunately, given the framework of this paper, those two goals are compatible. A significant additional advantage is that the usefulness of the translator aids component (levels one and two) will facilitate obtaining serious user feedback during the development of the machine translation component (level three).

Note

1. There are three groups doing work on machine-assisted translation in Provo, Utah, U.S.A. Two are commercial endeavors (Weidner and ALPS), and the third, the one described in this paper, is an academic research project at Brigham Young University. All three groups include researchers who participated in the development of ITS version one, yet all three are independent organizations.

References

Andreyewski, A. 1981. Translation Aids, Robots, and Automation, *META*, 26, 57–66.

Bar-Hillel, J. 1960. The Present Status of Automatic Translation of Languages, *Advances in Computers*, 1, 91–163. Reprinted in this collection.

Baudot, J., A. Clas, and I. Gross 1981. Un modèle de mini-banque de terminologie bilingue, *META*, 26, 315–331.

Boitet, C., P. Chatelin, and P. Daun Fraga. 1980. Present and Future Paradigms in the Automatized [sic] Translation of Natural Languages. In *COLING-80: Proceedings of the 8th International Conference on Computational Linguistics* (Tokyo), 430–436.

Brinkmann, K.-H. 1980. Terminology Data Banks as a Basis for High-Quality Translation. In *COLING-80: Proceedings of the 8th International Conference on Computational Linguistics* (Tokyo), 463.

Kay, M. 1980. The Proper Place of Men and Machines in Language Translation. Research Report CSL-80-11, Xerox Palo Alto Research Center. Reprinted in this collection.

Lippman, E. 1977. Computer Aids for the Human Translator. Report presented at the VID World Congress of FIT, Montreal.

Melby, A. K., M. R. Smith, and J. Peterson. 1980. ITS: Interactive Translation System. In *COLING-80: Proceedings of the 8th International Conference on Computational Linguistics* (Tokyo), 424–429.

Melby, A. K. 1981a. Linguistics and Machine Translation. In J. Copeland and P. Davis (eds.), *The Seventh LACUS Forum 1980*. Columbia, SC: Hornbeam Press.

Melby, A. K. 1981b. A Suggestion Box Translator Aid. In *Proceedings of the Annual Symposium of the Deseret Language and Linguistic Society*. Provo, UT: Brigham Young University.

Seppänen, J. 1979. *Pragmatic Aspects of Man/Computer Dialogues*, Research Report No. 12, Helsinki University of Technology Computer Center.

Silverstein, V. 1981. Letter to the Editor, *ATA Chronicle* (November 1981).

This page intentionally left blank

30
EUROTRA: Computational Techniques

Rod Johnson, Maghi King, and Louis des Tombe

Computational Techniques

In the present state of the art, the problem of machine translation is not fully understood. In some subdomains (e.g., English syntax, English–French lexical equivalences) we have a good deal of experience, a rich theoretical literature, and, hence, the confidence to predict in some detail the behaviour of the program to do the job. In other areas (the synthesis of Greek texts, mapping Italian representations to equivalent Danish text representations), we have virtually no experience and can only make informed guesses about the "right" way to do the job by computer. In the worst case we are still (at the time of writing at least) hopelessly at a loss when it comes to characterizing precisely what is preserved in translation if more than two languages are involved. In other words, we do not have, as yet, anything like a complete theory of multilingual machine translation. We have argued elsewhere, and at some length (Johnson et al. 1984, Johnson and Rosner 1987) that it is in the nature of problem-oriented software to embody some theory of the problem domain, and we shall not repeat the detailed arguments here.

We simply restate our view that no existing solution to the question of finding an appropriate problem-oriented programming language for machine translation seems to us to be acceptable for EUROTRA. These solutions fall roughly into three categories:

A. Assume some theory and implement it directly; this approach is fairly rare, but seems inherent, for instance, in Jan Landsbergen's Rosetta project (Landsbergen 1987).

B. Use an existing programming language, perhaps extended by a library of purposely built macros, subroutines or functions (depending on persuasion): examples of this approach are the IBM macro assembler in SYSTRAN (Bruderer 1978:100) and FORTRAN in SUSY (Maas 1984).

C. Invent a new programming language, embodying a very weak, low-level theory of machine translation usually based on explicit tree-to-tree mappings as in ROBRA (Boitet and Nedobejkine 1981), Q-systems (TAUM 1973), and GRADE (Tsujii 1983). The underlying thesis is normally sufficiently weak, in such cases, to allow the claim that the language has universal or near-universal application for all or most of machine translation tasks.

We have not adopted (A) because there is not sufficient practical evidence of a single theory that encompasses translations between all pairs of the Community languages. We reject (B) on the grounds that ordinary programming languages are just too unconstrained to be reliably handled by a large, loosely linked community of users, many of whom are unskilled in their use; and they obscure some of the true issues of linguistic knowledge representation and use in the detail of managing a von Neumann machine (or lambda calculus or Horn clauses or what have you). The last option, (C) is more interesting. In principle, we reject (C) also, although in the short term we have adopted a form of it for reasons of expediency, as we explain below. We are sceptical about any kind of universal programming language for machine translation, because we believe that the tasks involved in machine translation are essentially heterogeneous in nature. If we are constrained to use the same language to describe syntactic parsing, "semantic" interpretation, lexical and structural transfer, resolution of structural and lexical ambiguities, in and between seven different languages, it follows that either all of these are comparable or that the language of description gives us very little help in saying what we want to say.

To give a very simple example, suppose we have a strategy for parsing English that uses phrase structure recognition to construct a network of syntactic relations (subject, object, etc.) and then maps these relations to case relations like agent, patient, etc. In the homogeneous view of the world, we might have to write something like:

```
given A+B where cat(A)=NP and cat(B)=VP
build C(A+B) setting cat(C)=S
```

and

```
given A+B where cat(A)=V and cat(B)=NP
build C(A+B) where cat(C)= VP
```

followed in a later process by:

```
given A(B+C(X*+D+Y*+E+Z*)
   where cat(A)=S and cat(B)=NP and cat(C)=VP
   and cat(D)=V and cat (E)=NP
build P(Q+R+X*+Y*+Z*)
   where srel(P)=pred and srel(Q)=subj and
   srel(Q)=obj
   and lex(P)=lex(D) and lex(Q) =lex(B) and
   lex(R)= lex(E)
   and semf(P)=semf(D) and semf(Q)=semf(B) and
   semf(R)=semf(E)
```

/* **semf** stands for "semantic feature", X^*, Y^*, Z^* are intended to stand for variables over sequences of trees */

followed again later by

```
given A(B+C)
   where srel(A)=pred and srel(B)=subj and
   srel(C)=obj and action-process in semf(A)
   and animate in semf(B)
build A(B+C) adding case(A)=pred and
case(B)=agent
   and case(C)=patient
```

While the above notation is very informal, it is worth noting the very arbitrary semantics that underlie it. For example, there are clearly conventions about the use of identical variable names on the left- and right-hand side of rules; in some cases, right-hand nodes may be understood as copies of corresponding nodes on the left (indicated by the use of *where*), in others they may be interpreted as identified with their left-hand counterparts (indicated by *adding*). The arbitrariness is not accidental; indeed, since the linguistic theory underlying the notation is so weak, the meaning of the notation cannot but be arbitrary to the user. Their arbitrariness, however, is not the biggest defect of notations of this kind. Where they really fail is in being intolerably cluttered, since the user is forced to be explicit about every detail of the operations, precisely because in the absence of any strong linguistic theory, none of the responsibility for details can be left to the machine.

Consider now the same statements in a more perspicuous notation:

$S \rightarrow NP[^i \text{ SUBJ} =^*]\ VP[^i =^*]$
$VP \rightarrow V[^i =^*]\ NP[^i \text{ OBJ} =^*]$

and elsewhere (in the lexicon perhaps),

if action-process in semf(PRED) and animate in semf(SUBJ)
then [SUBJ → AGENT, OBJ → PATIENT]

Again the notation is informal, but not totally arbitrary (the debt to Lexical Functional Grammar (Kaplan and Bresnan 1982) is obvious). What is significant, though, is not so much the syntax of the notation as its semantics. Because we have a theory of parsing, we can include in the user's machine a large chunk of the meaning of what it is to parse within that theory. As a result, the user is left with a much clearer view of the task in hand: to provide the details of specific cases within the theory.

The ideal goal of the EUROTRA software design should be to provide just such a theory-sensitive system for machine translation. Unfortunately, and we have made the point many times here, we just do not have sufficient knowledge of the domain to provide the necessary theoretical input, and the problem is magnified in the special circumstances of EUROTRA.

What we have therefore built is an environment in which new theories and/or sub-theories of machine translation can be implemented very rapidly on an experimental basis. The environment consists essentially of four parts, not including the usual editing and debugging facilities. Two of the parts are quite standard: a compiler compiler, which we use to write compilers for the languages of a new theory; and a kernel interpreter that runs the outputs from the compiler. What is interesting is that we contrive to make the process of compilation as much as possible a purely syntactic one, mapping statements in the user language into a simple tuple language. Statements in the tuple language are not, however, executable directly by the kernel interpreter, since they contain as yet uninterpreted symbols. The interpretation of the symbols is given by external definitions, which are of two types: control definitions and data definitions. As the names suggest, data definitions are essentially instructions to a pattern matcher which acts as a slave to the main interpreter; control definitions define how and when calls to the pattern matcher are made. By judicious choice of the definition languages we are able to use these external definitions in two ways—to make rapid implementation of new theories, and to serve directly as specifications for a more efficient implementation, should the user agree after experimen-

tation to include a new theory in the system. A more detailed description can be found in Johnson et al. (1984).

This device is already proving very effective in allowing users to try out new ideas. More important, it frees us from the dangers of committing the user community too early to a small number of particular strategies, which may turn out to be unsuitable in the medium term, without making ultimate commitment impossible by imposing monolithic homogeneity from the start.

Nonetheless, we clearly need to make some decisions now, however provisional, so that we can get started. The remainder of this section describes the first user-language prototype implementation which is being handed over to users for preliminary experimentation.

All our software prototyping has been done under UNIX,[1] both for reasons of easy portability and because of the rich set of available software tools. The original prototype was developed on a VAX-11/780 under bsd version 4.2, and successful ports have been made to a bsd version 4.1 on a VAX-750 and to a Dual Systems 83/20 running Unisoft Version 7. We are about to attempt a port to a Sun Workstation and anticipate no serious difficulty.

It should be noted that our decision to adopt UNIX as a software prototyping environment (and therefore necessarily as a linguistic prototyping environment in the short term) does not necessarily of itself commit the anticipated industrial implementation to any particular hardware/software combination. The main purpose of our own software prototypes is to help us derive more reliable specifications for the industrial implementation, and to provide temporary short term support for linguistic experimentation.

The First User-language Prototype

Processes

The overriding design criterion we have followed is that of modular construction. Not only is this generally desirable, it is virtually essential given the organizational framework of EUROTRA. The basic unit of a user "program" is called a process. Since we want it to be possible for users to test parts of a system independently of others, and indeed to combine parts together in a reliable way, we have been particularly careful to provide ways of limiting or even excluding the propagation of unexpected side effects between processes. We achieve this by defining a process as a quintuple

```
process=[name, expectation, focus, body, goal]
```

The name is just a symbol used to identify the process. The expectation and the goal are pattern descriptions that serve a number of desirable functions. The most important of these is to guarantee that the domain and range of the process can be known when the process is defined. They achieve this by acting as filters over the currently active data configuration. A process may only operate on data that satisfy the expectation; correspondingly, only data that satisfy the goal are allowed to be output from the process. Operationally what happens is: the system attempts to apply the process by matching the expectation against the currently active data set; the process is invoked only if a match is found, in which case the process is applied in parallel to all data subsets that match; on termination (we assume that the process terminates) all results are matched against the goal; in all, and only, the cases where the match succeeds, the new results are added to the active data set, and the system proceeds to the next task.

The focus gives a way of narrowing down application to a subset of the data set yielded by the expectation; this is necessary, for example, when a process invokes itself recursively.

The process body may be either primitive or non-primitive. Processes with primitive bodies are also called grammars, and we shall return to them later. Non-primitive bodies consist of expressions over the names of processes, where the meaning of the expression can be varied by external definitions. In the current version, we allow regular expressions over processes, interpreting the concatenation operator as sequential application, the union operator as parallel application and the closure or star operator as all paths combinatorial application. The principle underlying this general scheme of controlling pattern directed invocation via a formal control language owes much to the work of Georgeff (1982). Thus, in the body of a process, a user might write

```
body
   p1,p2, (p3 | p4)
```

with the meaning "apply p1, then p2, then p3 and p4 in parallel." Our current compiler is defined to translate this into the tuple

```
[sequence, p1,p2, [parallel, p3,p4]]
```

And, in our control definition language (we currently use FP, Backus 1977), the definition of *apply* includes:

apply = atom → execute;
eq ∘ [1, 'sequence] → /apply ∘ tail;
eq ∘ [1, 'parallel] → apply ∘ tail

where, with some simplification

execute = integrate ∘ filter-goal ∘
apply ∘ filter-expectation.

It should be emphasised that the user is only concerned with writing (and understanding!) statements like

```
body p1,p2, (p3 | P4)
```

Grammars

The process interpreter continues to try to apply processes until it bottoms out at grammars (processes whose body is a primitive). The structure of a primitive depends on the theory it implements: thus a general rewrite primitive will be organized—and defined—differently from a dictionary primitive, which in its turn will differ from a transfer primitive, and so on. We currently have very few primitives, since the system is still in an experimental stage. The most important is a non-deterministic tree transducer, implementing a general rewrite system, which does not differ in any interesting way from Colmerauer's (1971) Q-system or Kay's (1967) powerful parser. Its main purpose is to provide users with a very (excessively) powerful tool for experimentation, and to provide fall-back for those cases where there is no adequate computational linguistic theory. We also have an analysis dictionary (a device that maps strings to nodes with complex collections of attributes and features) and a phrase structure parser. We are about to start on a transfer device and, as a more searching test of the capabilities of the basic tools, an implementation of a multilevel parser inspired by LFG. Once the basic tools were built, we found it very easy to build prototype implementations quickly. For example, the general rewrite system took about two man-months. The first dictionary implementation took less than a man-week. We expect that the transfer device will take around two to three weeks; the multilevel parser will almost certainly take longer—perhaps a month to six weeks.

Data Structure

In our system, there is no data "structure" as such. The same effect is achieved through interaction between a pattern matcher and a database of primitive objects called nodes. The behaviour of the pattern matcher is defined externally through statements in a data definition language, much in the same way as the meaning of system control constructs is defined in FP. At the present time, we are using Prolog to supply both the data base manager and the definition language. This is not totally satisfactory, and we expect to have a more appropriate "in-house" data definition language shortly. To give a flavour of our data definitions, we give a single example of the definition and use of a tree, in pseudo-Prolog.

First we define some basic relations, using built-in higher-order relations:

```
antisymetric(dom)
intransitive(dom)
irreflexive(dom)
$dom(x,x)
$dom(x,y) :-
   dom(x,z),
   $dom(z,y)  /* reflexive transitive closure
   */
tree(R,x) :-
   $dom(R,x)    /* tree x with root R */
```

If the notation #x in the user program means "bind x to a tree," then we define our compiler to translate #x to [tree x]. The control interpreter simply performs elementary syntactic manipulation on data requests and passes them directly to the data manager, [tree x] is transformed to tree (-,- x). Repeated calls to the data manager will yield all possible trees x in the currently active data set.

Disambiguation

The system potentially has a number of ways of dealing with ambiguity. Which ones are used depends on the extent to which disambiguation strategy is embedded into an implemented theory.

The simplest device is an extension of the use of goals to allow the user to supply an ordered list of goal descriptions. The system simply continues to try to match goals, in order, until it finds one which succeeds. The output from that goal is the result of the process. This rather cumbersome device is actually quite useful, for example in constructing elementary preference strategies painlessly. It is, however, not particularly subtle.

More interesting are strategies that exploit the inherent parallelism of the system—defined, for example through the (apply to all) functional of FP. Normally, the results of a parallel process application are all added to the current data set "in the same place." We could, however implement a primitive that allows the user to state criteria for selection be-

tween competing representations, and to exclude less favoured ones on the basis of linguistically motivated judgments. This would only be sensible, however, if the user were able to formulate such judgments in a general way.

Finally, we also have the option of implementing a relation **alt** (for alternative) directly in the data definitions (we have, in fact, done a simulation of a chart parser in this way). The problem here is that an alt relation between nodes is easy to handle, but an induced alternative relation between sets of nodes is not, unless the process that constructs it is very well behaved (for example, only building alternatives between simple constructs like trees). We do not know of any practical method of guaranteeing that such a relation can be maintained in a system which can perform transformations of arbitrary complexity.

Efficiency

The system we have described here is not particularly efficient—indeed it can be dramatically inefficient when presented with only moderately large and complex computations to perform. We are not (yet) unduly concerned by this inefficiency, for two reasons. First, we are still at the experimental stage where correctness is still more important than speed; there are no plans for an industrial implementation before 1988. Second, the experimental device we have described here has two equally important functions: the first is indeed to permit us to generate implementations of new theories rapidly for experimentation in the field; the second is to provide the basis for a formal specification of the semantics of that theory. If we can construct prototypes using precise definition languages, with the benign side effect that the same prototypes perform tolerably well for experimental purposes, we can be confident that an optimized implementation derived from the same specifications has a good chance of being both correct and operationally efficient.

Note

1. Trademark of AT&T Bell Laboratories.

References

Backus, J. 1977. Can Programming be Liberated from the von Neumann Style? *Communications of the ACM*, 21, 613–641.

Boitet, C., and N. Nedobejkine. 1981. Russian-French Machine Translation at Grenoble: A General Software Used for Implementing a Particular Linguistic Strategy. *Linguistics*, 19, 199–271.

Bruderer, H. E. 1978. *Handbuch der maschinellen und maschinenuntersützten Sprachübersetzung*. München: Verlag Dokumentation.

Colmerauer, A. 1971. *Les systèmes-Q, ou un formalisme pour analyser et synthétiser des phrases sur ordinateur*. Publication interne No. 43, Groupe TAUM, Université de Montréal.

Georgeff, M. 1982. Procedural Control in Production Systems. *Artificial Intelligence*, 18, 175–201.

Johnson, R. L., S. Krauwer, M. Rosner, and G. B. Varile. 1984. The Design of the Kernel Architecture of the EUROTRA System. *10th International Conference on Computational Linguistics, 22nd Annual Meeting of the Association for Computational Linguistics: Proceedings of COLING-84* (Stanford, California), 226–235.

Johnson, R., and M. Rosner. 1987. Machine Translation and Software Tools. In M. King (ed.), *Machine Translation Today*. Edinburgh: Edinburgh University Press, 154–167.

Kaplan, R. M., and J. Bresnan. 1982. Lexical Functional Grammar: A Formal System for Grammatical Representation. In J. Bresnan (ed.), *The Mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press, 173–281.

Kay, M. 1967. Experiments with a Powerful Parser. *2ème Conférence International sur le Traitement Automatique des Langues* (Grenoble).

Landsbergen, J. 1987. Isomorphic Grammars and their Use in the Rosetta Translation System. In M. King (ed.), *Machine Translation Today*. Edinburgh: Edinburgh University Press, 351–372.

Maas, H. D. 1984. The MT System SUSY. In M. King (ed.), *Machine Translation Today*. Edinburgh: Edinburgh University Press, 209–246.

TAUM 1973. Le système de traduction automatique de l'Université de Montréal (TAUM). *Meta*, 18, 227–289.

Tsujii, J-I. 1983. Technical Outlines of Japanese National MT Project. Paper given at the Joint EUROTRA-Japanese Workshop, Brussels.

This page intentionally left blank

31
A Framework of a Mechanical Translation between Japanese and English by Analogy Principle

Makoto Nagao

Prototypical Consideration

Let us reflect on the mechanism of human translation of elementary sentences at the beginning of foreign language learning. A student memorizes elementary English sentences with the corresponding Japanese sentences. The first stage is completely a drill of memorizing lots of similar sentences and words in English, and the corresponding Japanese. Here we have no translation theory at all to give to the student. He has to get the translation mechanism through his own instincts. He has to compare several different English sentences with the corresponding Japanese. He has to guess, make inferences about the structure of sentences from a lot of examples.

Along the same lines as this learning process, we shall start the consideration of our machine translation system, by giving lots of example sentences with their corresponding translations. The system must be able to recognize the similarity and the difference of the given example sentences. Initially a pair of sentences is given, a simple English sentence and the corresponding Japanese sentence. The next step is to give another pair of sentences (English and Japanese), which is different from the first only by one word (figure 31.1).

This word replacement operation is done one word at a time in the subject, object, and complement positions of a sentence with lots of different words. For each replacement one must give the information to the system of whether the sentence is acceptable or non-acceptable. Then the system will obtain at least the following information from this experiment:

(a) Certain facts about the structure of a sentence;

(b) Correspondence between English and Japanese words.

Results indicate that we can formulate a word dictionary between English and Japanese, and a set of noun groups by sentential context. If this experiment is done for different kinds of verbs the noun grouping will become much more fine and complex, and more reliable. Then certain kinds of relations will be established between word groups in a very complicated network structure. A noun may belong to several different groups with many different relations to other nouns. This is a kind of extensional representation of word meanings.

The same experiment can be done to verbs by replacing a verb in the same contextual environment. However, this is not so easy as noun replacement, because each verb has certain specific features as to the sentential structure, and no good grouping of verbs can be expected. So the sentential structure abstraction is done for each verb, and the structures are memorised in the verb dictionary entry for individual verb basis in such forms as (1).

(1) $\mathrm{S} \cdot \mathrm{verb} \cdot \mathrm{O} \cdot \mathrm{C} \Leftrightarrow \mathrm{S}'\ wa \cdot \mathrm{O}'\ wo \cdot \mathrm{C}'\ ni \cdot \mathrm{verb}'$,
$\mathrm{S}, \mathrm{S}' \in w_{\mathrm{X}}, \mathrm{O}, \mathrm{O}' \in w_{\mathrm{Y}}, \mathrm{C}, \mathrm{C}' \in w_{\mathrm{Z}}$
where $w_{\mathrm{X}}, w_{\mathrm{Y}}, w_{\mathrm{Z}}$ are semantic groups of words $\mathrm{X}, \mathrm{Y}, \mathrm{Z}$.

This is a procedure of finding the case frames for each verb mechanically. But to get a good and reliable result we have to have a huge amount of sample sentences which are carefully prepared. To distinguish word usages of similar nature, we sometimes have to prepare near-miss sentences. The data preparation of this kind is very difficult, and the speed of learning of the linguistic structures by the system is very slow.

A Modified Approach

To improve this simple language learning process, we can think of the utilization of ordinary word dictionaries and thesauri. In an ordinary word dictionary a verb has, in the explanation part, typical usages of the verb in example sentences rather than grammatical explanations. That is, typical sentential structures which the verb is governed by are given as examples. These dictionary examples give us plenty of information as to the sentential structures which the verb is governed by. Man is guided by these examples, makes inferences, and generates varieties of sentences.

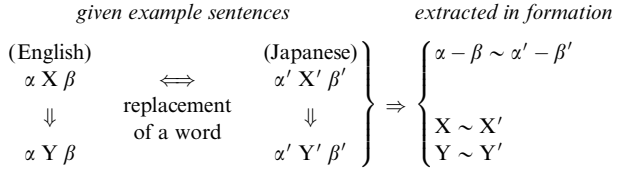


Figure 31.1

We want to incorporate this human process into our mechanical translation system. And for this purpose we need varieties of knowledge in our system. The knowledge the machine can utilize at the moment, however, is an ordinary word dictionary and thesaurus, which is of course not comparable to the human knowledge about the word and the sentences. A thesaurus is a system of word groupings of similar nature. It has information about synonyms, antonyms, upper and lower concept relations, part-whole relations and so on. The thesauri available at present are all very old, and they are not satisfactory from our standpoint, but we can use them properly.

The most important function in the utilization of example sentences in an ordinary dictionary is how to find out the similarity of the given input sentence and an example sentence, which can be a guide for the translation of the input sentence. First the global sytactic similarity between the input and example sentences must be checked. Then the replaceability of the corresponding words is tested by tracing the thesaurus relations. If the replaceability for every word is sufficiently sure, then the translation sentence of the example sentence is changed by replacing the words by the translation words of the input sentence. In this way the translation can be obtained.

For example, we are given an example sentence (2) for the verb *eat* from an English–Japanese dictionary, and its translation as sentence (3). Suppose sentence (4) is given for translation.

(2) A man eats vegetables.

(3) *hito-wa yasai-wo taberu.*
(man) (vegetable) (eat)

(4) He eats potatoes.

The system checks the replaceability (∼) of the words (5) by tracing the synonym and upper/lower concept relations in a thesaurus. Because these are similar word pairs, the system determines that the translated example (3) can be used for the translation of (4). From the dictionary the translation of the words (5) is (6) in the table, and the replaced result is (7) which is a good translation of the sentence (4).

(5) man ∼ he
vegetable ∼ potato

(6) *hito* (man) ∼ *kare* (he)
yasai (vegetable) ∼ *jagaimo* (potato)

(7) *kare-wa jagaimo-wo taberu.*

When sentence (8) is given, the similarity check (9) fails in the thesaurus, and no translation comes out.

(8) Acid eats metal.

(9) acid ∼ man
metal ∼ vegetable

If (8) is an example sentence in the dictionary entry for *eat*, and has the Japanese translation (10), then the input sentence (11) can be translated as (12).

(10) *san-wa kinzoku-wo okasu.*
(acid) (metal) $\begin{pmatrix} \text{eat} \\ \text{invade} \\ \text{attack} \end{pmatrix}$

(11) Sulphuric acid eats metal.

(12) *ryūsan-wa tetsu-wo okasu.*

The important point in this process is the recognition of the similarity between the input sentence and an example sentence in a dictionary. This completely depends on the structure of the thesaurus. Typical examples of *yabureru* ('be defeated', or 'be broken') are sentences (13) and (15), and the corresponding translations (14) [sic] and (16).

(13) *kare-wa senkyo-ni yabureta*
(he) (election) (be defeated)

(14) He was defeated by the election.

(15) *kamibukuro-wa omomi-de yabureta*
(paper bag) (weight) (be broken)

(16) The paper bag was broken by the weight.

Suppose we are given a sentence (17). To know which usage of *yabureru* fits this sentence, we check the words *president* and *vote* in a thesaurus, and find out the relations (18). We can determine from this information that (17) is more related to (13) than to (15), and the translation is obtained as (19).

(17) *daitōryō-wa tōhyō-ni yabureta.*
(president) (vote)

(18) *daitoryo* (president) ∼ *hito* (man)
tōhyō (vote) ∼ *senkyo* (election)

(19) The president was defeated by the vote.

[...]

Machine Translation by Analogy

Our fundamental ideas about the translation are:

(a) Man does not translate a simple sentence by doing deep linguistic analysis, rather,

(b) Man does the translation, first, by properly decomposing an input sentence into certain fragmental phrases (very often, into case frame units), then, by translating these fragmental phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference, which is illustrated above.

European languages have a certain common basis among them, and the mutual translation between these languages will be possible without great structural changes in sentential expressions. But translation between two languages which are totally different, like English and Japanese, has a lot of difficult problems. Sometimes the same contents are expressed by completely different sentential structures, and there is no good structural correspondence between each part of the sentences of the two languages.

For example, a Japanese sentence (20) corresponds to such different English sentences as (21)–(24). Another example is (25), which will literally correspond to such sentences as (26)–(28). But, it simply means (29).

(20) *zanen* *nagara* *ashita-wa*

(regret / disappointment) (though / inspite of / while / with.) (tomorrow)

ike *masen.*

(go / visit / attend) (not)

(21) To my regret I cannot go tomorrow.

(22) I am sorry I cannot visit tomorrow.

(23) It is a pity that I cannot go tomorrow.

(24) Sorry, tomorrow I will not be available.

(25) *kokusaiseiji* *no* *koto* *nitsuite*

(international (of) politics) (matter / thing / affair / situation / event) (about / of / on / with)

kaita *hon.*

(write / draw) (book / volume / work)

(26) a book in which the affairs of international politics is written

(27) a book in which (someone) wrote about the events of international politics

(28) a book written about the events of international politics

(29) a book on international politics

A translation of this kind cannot be achieved by a mere detailed syntactic analysis of the original sentence. If we pick up each word and look for the corresponding word in the translation, the synthesis of the target language sentence becomes almost impossible. The choice of the proper translation from many candidates for a source language word is also very difficult without seeing the wider sentential context.

Therefore we adopted the method which may be called machine translation by example-guided inference, or machine translation by the analogy principle, and whose fundamental idea has been introduced above. One of the strong reasons for this approach has been that the detailed analysis of a source language sentence is of no use for translation between languages of completely different structure like English and Japanese. We have to see as wide a scope as possible in a sentence, and the translation must be from a block of words to a block of words. To realize this we have to store varieties of example sentences in the dictionary and to have a mechanism for finding analogical example sentences for the given one.

It is very important to point out that if we want to construct a system of learning, we have to be able to give the system the data which is not very much processed. In our system the augmentation of the knowledge is very simple and easy. It requires only the addition of new words and new usage examples and their translations. It does not require the information which is deeply analyzed and well arranged. Linguistic theories change rapidly to and fro, and

sometimes a model must by thrown away after a few years. On the contrary, language data and its usage do not change for a long time. We will rely on the primary data rather than analyzed data which may change sometimes because of changes in the theory.

A Practical Approach

The process of mechanical translation by analogy is again very time consuming in its primary structure. So we divide the process into a few substages and give all the available information we have to the system, in the initial system construction. The learning comes in only at the augmentation stage of the system, which is mainly the increase of example sentences and the improvement of the thesaurus.

The following substages have been distinguished in our Japanese–English translation system which is being constructed.

(a) Reduction of redundant expressions, and supplement of eliminated expressions in a Japanese input sentence, and getting an essential sentential structure. Sentence (30) has almost the same meaning as sentence (31).

(30) *nihongo-no honyaku-no baai-ni tsuite-wa*,
(Japanese) (translation) (case) (about)
muzukashii mondai-ga aru.
(difficult) (problem) (exist)

(31) *nihongo-no honyaku ni-wa muzukashii mondai-ga aru*.

(b) Analysis of sentential structure by case grammar. Phrase structure grammar is not suitable for the analysis of Japanese, because the word order in Japanese is almost free except that the final predicate verb comes at the end.

(c) Retrieval of target language words and example phrases which are stored in the word entries from the dictionary. The dictionary contains varieties of examples besides grammatical information, meaning and, for verbs, the case frames.

(d) Recognition of the similarity between the input sentential phrases and example phrases in the dictionary. The word thesaurus is used for the similarity finding.

(e) Choice of a global sentential form for translation. For example, sentence (32) has such translations as (33) and (34). These can only be derived from the examples for the word result in the dictionary.

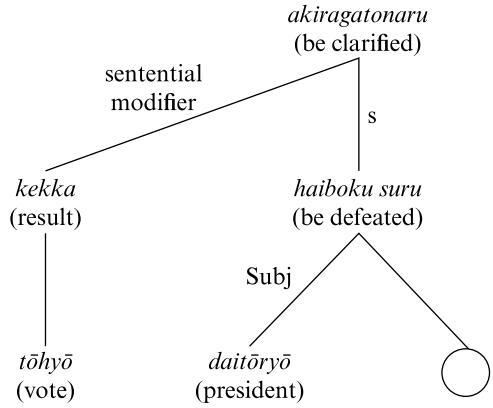


Figure 31.2

(32) *tōhyō no kekka daitōryō no haiboku-ga*
(vote) (of) (result) (president) (of) (defeat)
akira-gato natta.
(clear / evident) (become)
(see figure 31.2).

(33) As the result of the vote the defeat of the president became definite.

(34) The result of the vote revealed that the president was defeated.

(f) The choice of local phrase structure is determined by the requirements of the global sentential structure.

It is very difficult to clarify what factors contribute to the determination of the stages (e) and (f). These remain to be solved.

32
A Statistical Approach to Machine Translation

Peter F. Brown, John Cocke, Stephen A. Della Pietra,
Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty,
Robert L. Mercer, and Paul S. Roossin

Introduction

The field of machine translation is almost as old as the modern digital computer. In 1949 Warren Weaver suggested that the problem be attacked with statistical methods and ideas from information theory, an area which he, Claude Shannon, and others were developing at the time (Weaver 1949). Although researchers quickly abandoned this approach, advancing numerous theoretical objections, we believe that the true obstacles lay in the relative impotence of the available computers and the dearth of machine readable text from which to gather the statistics vital to such an attack. Today, computers are five orders of magnitude faster than they were in 1950 and have hundreds of millions of bytes of storage. Large, machine-readable corpora are readily available. Statistical methods have proven their value in automatic speech recognition (Bahl et al. 1983) and have recently been applied to lexicography (Sinclair 1985) and to natural language processing (Baker 1979; Ferguson 1980; Garside et al. 1987; Sampson 1986; Sharman et al. 1988). We feel that it is time to give them a chance in machine translation.

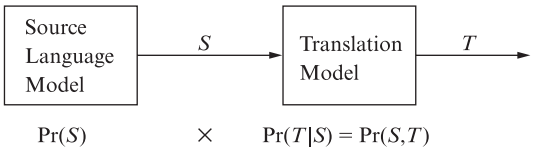
The job of a translator is to render in one language the meaning expressed by a passage of text in another language. This task is not always straightforward. For example, the translation of a word may depend on words quite far from it. Some English translators of Proust's seven-volume work *A la Recherche du Temps Perdu* have striven to make the first word of the first volume the same as the last word of the last volume because the French original begins and ends with the same word (Bernstein 1988). Thus, in its most highly developed form, translation involves a careful study of the original text and may even encompass a detailed analysis of the author's life and circumstances. We, of course, do not hope to reach these pinnacles of the translator's art.

In this paper, we consider only the translation of individual sentences. Usually, there are many acceptable translations of a particular sentence, the choice among them being largely a matter of taste. We take the view that every sentence in one language is a possible translation of any sentence in the other. We assign to every pair of sentences (S, T) a probability, $\Pr(T \mid S)$, to be interpreted as the probability that a translator will produce T in the target language when presented with S in the source language. We expect $\Pr(T \mid S)$ to be very small for pairs like (*Le matin je me brosse les dents* | *President Lincoln was a good lawyer*) and relatively large for pairs like (*Le président Lincoln était un bon avocat* | *President Lincoln was a good lawyer*). We view the problem of machine translation then as follows. Given a sentence T in the target language, we seek the sentence S from which the translator produced T. We know that our chance of error is minimized by choosing that sentence S that is most probable given T. Thus, we wish to choose S so as to maximize $\Pr(S \mid T)$. Using Bayes' theorem, we can write

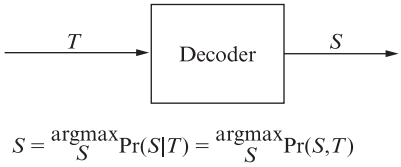
$$\Pr(S \mid T) = \frac{\Pr(S)\Pr(T \mid S)}{\Pr(T)}$$

The denominator on the right of this equation does not depend on S, and so it suffices to choose the S that maximizes the product $\Pr(S)\Pr(T \mid S)$. Call the first factor in this product the language model probability of S and the second factor the translation probability of T given S. Although the interaction of these two factors can be quite profound, it may help the reader to think of the translation probability as suggesting words from the source language that might have produced the words that we observe in the target sentence and to think of the language model probability as suggesting an order in which to place these source words.

Thus, as illustrated in figure 32.1, a statistical translation system requires a method for computing language model probabilities, a method for computing translation probabilities, and, finally, a method for searching among possible source sentences S for the one that gives the greatest value for $\Pr(S)\Pr(T \mid S)$.



A *Source Language Model* and a *Translation Model* furnish a joint probability distribution over source—target sentence pairs (*S*, *T*). The joint probability Pr(*S*, *T*) of the pair (*S*, *T*) is the product of the probability Pr(*S*) computed by the language model and the conditional probability Pr(*T*|*S*) computed by the translation model. The parameters of these models are estimated automatically from a large database of source—target sentence pairs using a statistical algorithm which optimizes, in an appropriate sense, the fit between the models and the data.



A *Decoder* performs the actual translation. Given a sentence T in the target language, the decoder chooses a viable translation by selecting that sentence *S* in the source language for which the probability Pr(*S*|*T*) is maximum.

Figure 32.1
A statistical machine translation system.

In the remainder of this paper we describe a simple version of such a system that we have implemented. In the next section we describe our language model for Pr(S), and in section 3 we describe our translation model for Pr($T \mid S$). In section 4 we describe our search procedure. In section 5 we explain how we estimate the parameters of our models from a large database of translated text. In section 6 we describe the results of two experiments we performed using these models. Finally, in section 7 we conclude with a discussion of some improvements that we intend to implement.

The Language Model

Given a word string, $s_1 s_2 \ldots s_n$, we can, without loss of generality, write

$$\Pr(s_1 s_2 \ldots s_n) = \Pr(s_1) \Pr(s_2 \mid s_1) \ldots \Pr(s_n \mid s_1 s_2 \ldots s_{n-1})$$

Thus, we can recast the language modeling problem as one of computing the probability of a single word given all of the words that precede it in a sentence. At any point in the sentence, we must know the probability of an object word, s_j, given a history, $s_1 s_2 \ldots s_{j-1}$. Because there are so many histories, we cannot simply treat each of these probabilities as a separate parameter. One way to reduce the number of parameters is to place each of the histories into an equivalence class in some way and then to allow the probability of an object word to depend on the history only through the equivalence class into which that history falls. In an n-gram model, two histories are equivalent if they agree in their final $n-1$ words. Thus, in a bigram model, two histories are equivalent if they end in the same word and in a trigram model, two histories are equivalent if they end in the same two words.

While n-gram models are linguistically simple minded, they have proven quite valuable in speech recognition and have the redeeming feature that they are easy to make and to use. We can see the power of a trigram model by applying it to something that we call bag translation from English into English. In bag translation we take a sentence, cut it up into words, place the words in a bag, and then try to recover the sentence given the bag. We use the n-gram model to rank different arrangements of the words in the bag. Thus, we treat an arrangement S as better than another arrangement S' if Pr(S) is greater than Pr(S'). We tried this scheme on a random sample of sentences. From a collection of 100 sentences, we considered the 38 sentences with fewer than 11 words each. We had to restrict the length of the sentences because the number of possible rearrangements grows exponentially with sentence length. We used a trigram language model that had been constructed for a speech recognition system. We were able to recover 24 (63%) of the sentences exactly. Sometimes, the sentence that we found to be most probable was not an exact reproduction of the original, but conveyed the same meaning. In other cases, the most probable sentence according to our model was just garbage. If we count as correct all of the sentences that retained the meaning of the original, then 32 (84%) of the 38 were correct. Some examples of the original sentences and the sentences recovered from the bags are shown in figure 32.2. We have no doubt that if we had been able to handle longer sentences, the results would have been worse and that probability of error grows rapidly with sentence length.

The Translation Model

For simple sentences, it is reasonable to think of the French translation of an English sentence as being generated from the English sentence word by word.

Exact reconstruction (24 of 38)
Please give me your response as soon as possible.
⇒ Please give me your response as soon as possible.

Reconstruction preserving meaning (8 of 38)
Now let me mention some of the disadvantages.
⇒ Let me mention some of the disadvantages now.

Garbage reconstruction (6 of 38)
In our organization research has two missions.
In our missions research organization has two.

Figure 32.2
Bag model examples.

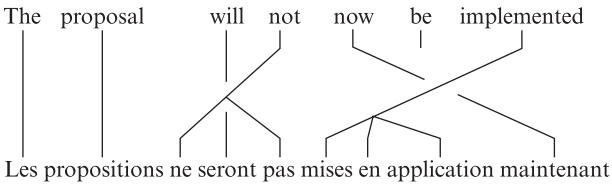


Figure 32.3
Alignment example.

Thus, in the sentence pair (*Jean aime Marie* | *John loves Mary*) we feel that *John* produces *Jean*, *loves* produces *aime*, and *Mary* produces *Marie*. We say that a word is *aligned* with the word that it produces. Thus *John* is aligned with *Jean* in the pair that we just discussed. Of course, not all pairs of sentences are as simple as this example. In the pair (*Jean n'aime personne* | *John loves nobody*), we can again align *John* with *Jean* and *loves* with *aime*, but now, *nobody* aligns with both *n*' and *personne*. Sometimes, words in the English sentence of the pair align with nothing in the French sentence, and similarly, occasionally words in the French member of the pair do not appear to go with any of the words in the English sentence. We refer to a picture such as that shown in figure 32.3 as an alignment. An alignment indicates the origin in the English sentence of each of the words in the French sentence. We call the number of French words that an English word produces in a given alignment its *fertility* in that alignment.

If we look at a number of pairs, we find that words near the beginning of the English sentence tend to align with words near the beginning of the French sentence and that words near the end of the English sentence tend to align with words near the end of the French sentence. But this is not always the case. Sometimes, a French word will appear quite far from the English word that produced it. We call this effect *distortion*. Distortions will, for example, allow adjectives to precede the nouns that they modify in English but to follow them in French.

It is convenient to introduce the following notation for alignments. We write the French sentence followed by the English sentence and enclose the pair in parentheses. We separate the two by a vertical bar. Following each of the English words, we give a parenthesized list of the positions of the words in the French sentence with which it is aligned. If an English word is aligned with no French words, then we omit the list. Thus (*Jean aime Marie* | *John*(*1*) *loves*(*2*) *Mary*(*3*)) is the simple alignment with which we began this discussion. In the alignment (*Le chien est battu par Jean* | *John*(*6*) *does beat*(*3*,*4*) *the*(*1*) *dog*(*2*)), *John* produces *Jean*, *does* produces nothing, *beat* produces *est battu*, *the* produces *Le*, *dog* produces *chien*, and *par* is not produced by any of the English words.

Rather than describe our translation model formally, we present it by working an example. To compute the probability of the alignment (*Le chien est battu par Jean* | *John*(*6*) *does beat*(*3*,*4*) *the*(*1*) *dog*(*2*)), begin by multiplying the probability that *John* has fertility 1 by $\Pr(\textit{Jean} \mid \textit{John})$. Then multiply by the probability that *does* has fertility 0. Next, multiply by the probability that *beat* has fertility 2 times $\Pr(\textit{est} \mid \textit{beat}) \Pr(\textit{battu} \mid \textit{beat})$, and so on. The word *par* is produced from a special English word which is denoted by $\langle \textit{null} \rangle$. The result is

$$\Pr(\textit{fertility} = 1 \mid \textit{John}) \times \Pr(\textit{Jean} \mid \textit{John}) \times$$

$$\Pr(\textit{fertility} = 0 \mid \textit{does}) \times$$

$$\Pr(\textit{fertility} = 2 \mid \textit{beat}) \times \Pr(\textit{est} \mid \textit{beat}) \Pr(\textit{battu} \mid \textit{beat}) \times$$

$$\Pr(\textit{fertility} = 1 \mid \textit{the}) \times \Pr(\textit{Le} \mid \textit{the}) \times$$

$$\Pr(\textit{fertility} = 1 \mid \textit{dog}) \times \Pr(\textit{chien} \mid \textit{dog}) \times$$

$$\Pr(\textit{fertility} = 1 \mid \langle \textit{null} \rangle) \times \Pr(\textit{par} \langle \textit{null} \rangle).$$

Finally, factor in the distortion probabilities. Our model for distortions is, at present, very simple. We assume that the position of the target word depends only on the length of the target sentence and the position of the source word. Therefore, a distortion probability has the form $\Pr(i \mid j, l)$ where i is a target position, j a source position, and l the target length.

In summary, the parameters of our translation model are a set of fertility probabilities $\Pr(n \mid e)$ for each English word e and for each fertility n from 0 to some moderate limit, in our case 25; a set of translation probabilities $\Pr(f \mid e)$, one for each element f of the French vocabulary and each member e of the English vocabulary; and a set of distortion proba-

bilities $\Pr(i \mid j, l)$ for each target position i, source position j, and target length l. We limit i, j, and l to the range 1 to 25.

Searching

In searching for the sentence S that maximizes $\Pr(S)\Pr(T \mid S)$, we face the difficulty that there are simply too many sentences to try. Instead, we must carry out a suboptimal search. We do so using a variant of the *stack search* that has worked so well in speech recognition (Bahl et al. 1983). In a stack search, we maintain a list of partial alignment hypotheses. Initially, this list contains only one entry corresponding to the hypothesis that the target sentence arose in some way from a sequence of source words that we do not know. In the alignment notation introduced earlier, this entry might be (*Jean aime Marie* | *) where the asterisk is a place holder for an unknown sequence of source words. The search proceeds by iterations, each of which extends some of the most promising entries on the list. An entry is extended by adding one or more additional words to its hypothesis. For example, we might extend the initial entry above to one or more of the following entries:

(*Jean aime Marie* | *John(1)**),

(*Jean aime Marie* | **loves(2)**),

(*Jean aime Marie* | **Mary(3)*),

(*Jean aime Marie* | *Jeans(1)**).

The search ends when there is a complete alignment on the list that is significantly more promising than any of the incomplete alignments. Sometimes, the sentence S' that is found in this way is not the same as the sentence S that a translator might have been working on. When S' itself is not an acceptable translation, then there is clearly a problem. If $\Pr(S')\Pr(T \mid S')$ is greater than $\Pr(S)\Pr(T \mid S)$, then the problem lies in our modeling of the language or of the translation process. If, however, $\Pr(S')\Pr(T \mid S')$ is less than $\Pr(S)\Pr(T \mid S)$, then our search has failed to find the most likely sentence. We call this latter type of failure a search error. In the case of a search error, we can be sure that our search procedure has failed to find the most probable source sentence, but we cannot be sure that were we to correct the search we would also correct the error. We might simply find an even more probable sentence that nonetheless is incorrect. Thus, while a search error is a clear indictment of the search procedure, it is not an acquittal of either the language model or the translation model.

Parameter Estimation

Both the language model and the translation model have many parameters that must be specified. To estimate these parameters accurately, we need a large quantity of data. For the parameters of the language model, we need only English text, which is available in computer-readable form from many sources; but for the parameters of the translation model, we need pairs of sentences that are translations of one another.

By law, the proceedings of the Canadian parliament are kept in both French and English. As members rise to address a question before the house or otherwise express themselves, their remarks are jotted down in whichever of the two languages is used. After the meeting adjourns, a collection of translators begins working to produce a complete set of the proceedings in both French and English. These proceedings are called Hansards, in remembrance of the publisher of the proceedings of the British parliament in the early 1800s. All of these proceedings are available in computer-readable form, and we have been able to obtain about 100 million words of English text and the corresponding French text from the Canadian government. Although the translations are not made sentence by sentence, we have been able to extract about three million pairs of sentences by using a statistical algorithm based on sentence length. Approximately 99% of these pairs are made up of sentences that are actually translations of one another.

In the experiments we describe later, we use a bigram language model. Thus, we have one parameter for every pair of words in the source language. We estimate these parameters from the counts of word pairs in a large sample of text from the English part of our Hansard data using a method described by Jelinek and Mercer (1980).

Earlier we discussed alignments of sentence pairs. If we had a collection of aligned pairs of sentences, then we could estimate the parameters of the translation model by counting, just as we do for the language model. However, we do not have alignments but only the unaligned pairs of sentences. This is exactly analogous to the situation in speech recognition where one has the script of a sentence and the time waveform corresponding to an utterance of it, but no indication of just what in the time waveform corresponds to what in the script. In speech recognition, this problem is attacked with the EM algorithm (Baum 1972;

Dempster et al. 1977). We have adapted this algorithm to our problem in translation. In brief, it works like this: given some initial estimate of the parameters, we can compute the probability of any particular alignment. We can then re-estimate the parameters by weighing each possible alignment according to its probability as determined by the initial guess of the parameters. Repeated iterations of this process lead to parameters that assign ever greater probability to the set of sentence pairs that we actually observe. This algorithm leads to a local maximum of the probability of the observed pairs as a function of the parameters of the model. There may be many such local maxima. The particular one at which we arrive will, in general, depend on the initial choice of parameters.

Two Pilot Experiments

In our first experiment, we test our ability to estimate parameters for the translation model. We chose as our English vocabulary the 9,000 most common words in the English part of the Hansard data, and as our French vocabulary the 9,000 most common French words. For the purposes of this experiment, we replaced all other words with either the *unknown English word* or the *unknown French word*, as appropriate. We applied the iterative algorithm discussed above in order to estimate some 81 million parameters from 40,000 pairs of sentences comprising a total of about 800,000 words in each language. The algorithm requires an initial guess of the parameters. We assumed that each of the 9,000 French words was equally probable as a translation of any of the 9,000 English words; we assumed that each of the fertilities from 0 to 25 was equally probable for each of the 9,000 English words; and finally, we assumed that each target position was equally probable given each source position and target length. Thus, our initial choices contained very little information about either French or English.

Figure 32.4 shows the translation and fertility probabilities we estimated for the English word *the*. We see that, according to the model, *the* translates most frequently into the French articles *le* and *la*. This is not surprising, of course, but we emphasize that it is determined completely automatically by the estimation process. In some sense, this correspondence is inherent in the sentence pairs themselves.

Figure 32.5 shows these probabilities for the English word *not*. As expected, the French word *pas* appears as a highly probable translation. Also, the fertility probabilities indicate that *not* translates most often into two French words, a situation consistent with the fact that negative French sentences contain the auxiliary word *ne* in addition to a primary negative word such as *pas* or *rien*.

For both of these words, we could easily have discovered the same information from a dictionary. In figure 32.6, we see the trained parameters for the English word *hear*. As we would expect, various forms of the French word *entendre* appear as possible translations, but the most probable translation is the French word *bravo*. When we look at the fertilities here, we see that the probability is about equally divided between fertility 0 and fertility 1. The reason for

English:	the		
French	**Probability**	**Fertility**	**Probability**
le	.610	1	.871
la	.178	0	.124
l'	.083	2	.004
les	.023		
ce	.013		
il	.012		
de	.009		
et	.007		
que	.007		

Figure 32.4
Probabilities for *the*.

English:	not		
French	**Probability**	**Fertility**	**Probability**
pas	.469	2	.758
ne	.460	0	.133
non	.024	1	.106
pas du tout	.003		
faux	.003		
plus	.002		
ce	.002		
que	.002		
jamais	.002		

Figure 32.5
Probabilities for *not*.

English:	hear		
French	**Probability**	**Fertility**	**Probability**
bravo	.992	0	.584
entendre	.005	1	.416
entendu	.002		
entends	.001		

Figure 32.6
Probabilities for *hear*.

this is that the English-speaking members of parliament express their approval by shouting *Hear*, *hear!*, while the French-speaking ones say *Bravo!* The translation model has learned that usually two *hear*s produce one *bravo* by having one of them produce the *bravo* and the other produce nothing.

A given pair of sentences has many possible alignments, since each target word can be aligned with any source word. A translation model will assign significant probability only to some of the possible alignments, and we can gain further insight about the model by examining the alignments that it considers most probable. We show one such alignment in figure 32.3. Observe that, quite reasonably, *not* is aligned with *ne* and *pas*, while *implemented* is aligned with the phrase *mises en application*. We can also see here a deficiency of the model since intuitively we feel that *will* and *be* act in concert to produce *seront* while the model aligns *will* with *seront* but aligns *be* with nothing.

In our second experiment, we used the statistical approach to translate from French to English. To have a manageable task, we limited the English vocabulary to the 1,000 most frequently used words in the English part of the Hansard corpus. We chose the French vocabulary to be the 1,700 most frequently used French words in translations of sentences that were completely covered by the 1,000-word English vocabulary. We estimated the 17 million parameters of the translation model from 117,000 pairs of sentences that were completely covered by both our French and English vocabularies. We estimated the parameters of the bigram language model from 570,000 sentences from the English part of the Hansard data. These sentences contain about 12 million words altogether and are not restricted to sentences completely covered by our vocabulary.

We used our search procedure to decode 73 new French sentences from elsewhere in the Hansard data. We assigned each of the resulting sentences a category according to the following criteria. If the decoded sentence was exactly the same as the actual Hansard translation, we assigned the sentence to the *exact* category. If it conveyed the same meaning as the Hansard translation but in slightly different words, we assigned it to the *alternate* category. If the decoded sentence was a legitimate translation of the French sentence but did not convey the same meaning as the Hansard translation, we assigned it to the *different* category. If it made sense as an English sentence but could not be interpreted as a translation of the French sentence, we assigned it to the *wrong* category. Finally, if the decoded sentence was grammatically deficient, we assigned it to the *ungrammatical* category. An example from each category is shown in figure 32.7, and our decoding results are summarized in figure 32.8.

Only 5% of the sentences fell into the exact category. However, we feel that a decoded sentence that is in any of the first three categories (exact, alternate, or different) represents a reasonable translation. By this criterion, the system performed successfully 48% of the time.

As an alternate measure of the system's performance, one of us corrected each of the sentences in the last three categories (different, wrong, and ungrammatical) to either the exact or the alternate category. Counting one stroke for each letter that must be deleted and one stroke for each letter that must be inserted, 776 strokes were needed to repair all of the decoded sentences. This compares with the 1,916 strokes required to generate all of the Hansard translations from scratch. Thus, to the extent that translation time can be equated with key strokes, the system reduces the work by about 60%.

Plans

There are many ways in which the simple models described in this paper can be improved. We expect some improvement from estimating the parameters on more data. For the experiments described above, we estimated the parameters of the models from only a small fraction of the data we have available: for the translation model, we used only about one percent of our data, and for the language model, only about ten percent.

We have serious problems in sentences in which the translation of certain source words depends on the translation of other source words. For example, the translation model produces *aller* from *to go* by producing *aller* from *go* and nothing from *to*. Intuitively we feel that *to go* functions as a unit to produce *aller*. While our model allows many target words to come from the same source word, it does not allow several source words to work together to produce a single target word. In the future, we hope to address the problem of identifying groups of words in the source language that function as a unit in translation. This may take the form of a probabilistic division of the source sentence into groups of words.

At present, we assume in our translation model that words are placed into the target sentence indepen-

Exact

	Ces ammendements sont certainement nécessaires.
Hansard:	These amendments are certainly necessary.
Decoded as:	These amendments are certainly necessary.

Alternate

	C'est pourtant très simple.
Hansard:	Yet it is very simple.
Decoded as:	It is still very simple.

Different

	J'ai reçu cette demande en effet.
Hansard:	Such a request was made.
Decoded as:	I have received this request in effect.

Wrong

	Permettez que je donne un exemple à la Chambre.
Hansard:	Let me give the House one example.
Decoded as:	Let me give an example in the House.

Ungrammatical

	Vous avez besoin de toute l'aide disponible.
Hansard:	You need all the help you can get.
Decoded as:	You need of the whole benefits available.

Figure 32.7
Translation examples.

Category	Number of sentences	Percent
Exact	4	5
Alternate	18	25
Different	13	18
Wrong	11	15
Ungrammatical	27	37
Total	*73*	

Figure 32.8
Translation results.

dently of one another. Clearly, a more realistic assumption must account for the fact that words form phrases in the target sentence that are translations of phrases in the source sentence and that the target words in these phrases will tend to stay together even if the phrase itself is moved around. We are working on a model in which the positions of the target words produced by a particular source word depend on the identity of the source word and on the positions of the target words produced by the previous source word.

We are preparing a trigram language model that we hope will substantially improve the performance of the system. A useful information-theoretic measure of the complexity of a language with respect to a model is the perplexity as defined by Bahl et al. (1983). With the bigram model that we are currently using, the source text for our 1,000-word translation task has a perplexity of about 78. With the trigram model that we are preparing, the perplexity of the source text is about 9. In addition to showing the strength of a trigram model relative to a bigram model, this also indicates that the 1,000-word task is very simple.

We treat words as unanalyzed wholes, recognizing no connection, for example, between *va*, *vais*, and *vont*, or between *tall*, *taller*, and *tallest*. As a result, we cannot improve our statistical characterization of *va*, say, by observation of sentences involving *vont*. We are working on morphologies for French and English so that we can profit from statistical regularities that our current word-based approach must overlook.

Finally, we treat the sentence as a structureless sequence of words. Sharman et al. discuss a method for deriving a probabilistic phrase structure grammar automatically from a sample of parsed sentences (1988). We hope to apply their method to construct grammars for both French and English and to base future translation models on the grammatical constructs thus defined.

References

Bahl, L. R., F. Jelinek, and R. L. Mercer. 1983. Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-5(2), 179–190.

Baker, J. K. 1979. Stochastic Modeling for Automatic Speech Understanding. In R. A. Reddy (ed.), *Speech Recognition*. New York: Academic Press.

Baum, L. E. 1972. An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of a Markov Process. *Inequalities*, 3, 1–8.

Bernstein, R. 1988. Howard's Way. *New York Times Magazine*, 138 (47639), 40–44, 74, 92.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39 (B), 1–38.

Ferguson, J. D. 1980. Hidden Markov Analysis: An Introduction. In J. D. Ferguson (ed.), *Hidden Markov Models for Speech*. Princeton, NJ: IDA-CRD.

Garside, R. G., G. N. Leech, and G. R. Sampson. 1987. *The Computational Analysis of English: A Corpus-Based Approach*. New York: Longman.

Jelinek, F., and R. L. Mercer. 1980. Interpolated Estimation of Markov Source Parameters from Sparse Data. *Proceedings of the Workshop on Pattern Recognition in Practice*. Amsterdam: North-Holland.

Sampson, G. R. 1986. A Stochastic Approach to Parsing. *Proceedings of the 11th International Conference on Computational Linguistics* (Bonn), 151–155.

Sharman, R. A., F. Jelinek, and R. L. Mercer. 1988. Generating a Grammar for Statistical Training. *Proceedings of the IBM Conference on Natural Language Processing* (Thornwood, NY).

Sinclair, J. M. 1985. Lexicographic Evidence. In R. Ilson (ed.), *Dictionaries, Lexicography and Language Learning*, New York: Pergamon Press.

Weaver, W. 1955. Translation (1949). In *Machine Translation of Languages*. Cambridge, MA: MIT Press. Reprinted in this volume.

33
Automatic Speech Translation at ATR

Tsuyoshi Morimoto and Akira Kurematsu

Introduction

Since Graham Bell first invented the telephone in 1876, it has become an indispensable means for communications. We can easily communicate with others domestically as well as internationally. However, another great barrier has not been overcome yet: communications between people speaking different languages.

An interpreting telephone system, or a speech translation system, will solve this problem which has been annoying human beings from the beginning of their history. The first effort was made by NEC; they demonstrated a system in Telecom'83 held in Geneva. In 1987, British Telecom Research Laboratories implemented an experimental system which was based on fixed phrase translation (Steer 1987). At Carnegie Mellon University (CMU), a speech translation system was developed for the doctor–patient domain in 1988 (Saito and Tomita 1988). These systems were small and simple, but showed the possibility of speech translation.

In Japan, the ATR Interpreting Telephony Research project, started in 1986 and terminated in March 1993, focused on basic research for speech translation and obtained fruitful results. Following that project, an Interpreting Telecommunication project was recently initiated.

This paper reports the technologies attained in the preceding project, and also describes the objectives of the new project.

Current Status of Basic Technologies for Speech Translation

In principle, three componential technologies are essential: speech recognition, language translation and speech synthesis. Furthermore, techniques concerning how to integrate speech recognition and language analysis are important. In this section, the technologies attained so far are described.

Speech Recognition

Basically, two kinds of models are necessary for speech recognition: a phonetic model and a language model. For phonetic modeling, a Hidden Markov Model (HMM) approach was employed. A phone is apt to be acoustically affected by preceding and/or succeeding phones, so hundreds of allophone models are generated automatically from a huge speech database by use of the "successively-state-splitting" (SSS) algorithm (Takami and Sagayama 1992). For the language model, a general context free grammar (CFG) was used. Compared to other conventional language models such as bigram or trigram, it is superior in extendability and maintainability. A new mechanism, a predictive LR parsing mechanism which is an extension of the generalized LR parsing algorithm, combines these two models dynamically and recognizes input continuous speech (Kita et al. 1989). In this method, CFG rules are compiled and converted to an LR table. The parser refers to the table and predicts the next possible phones, then verifies their existence in the input speech by comparison with corresponding HMMs (figure 33.1). As will be shown later in this section, this method attains a very high recognition rate.

As for non-specific user's speech, a speaker adaptation approach was adopted. By introducing the "vector field smoothing" (VFS) algorithm (Ohkura et al. 1992), only about ten words are sufficient to adapt to a new speaker's speech.

Integration of Speech Recognition and Language Analysis

The system accepts speech uttered phrase by phrase (Japanese *bunsetsu*) so that the speech is uttered clearly. To treat such utterances, Japanese phrasal grammar rules are defined in the speech recognizer. In addition to them, sentential level (inter-phrasal) grammar rules are defined, which are used by the sentence recognition controller. It controls inter-phrase level parsing, and, coping with the phrase recognizer, recognizes input sentences as a whole rather

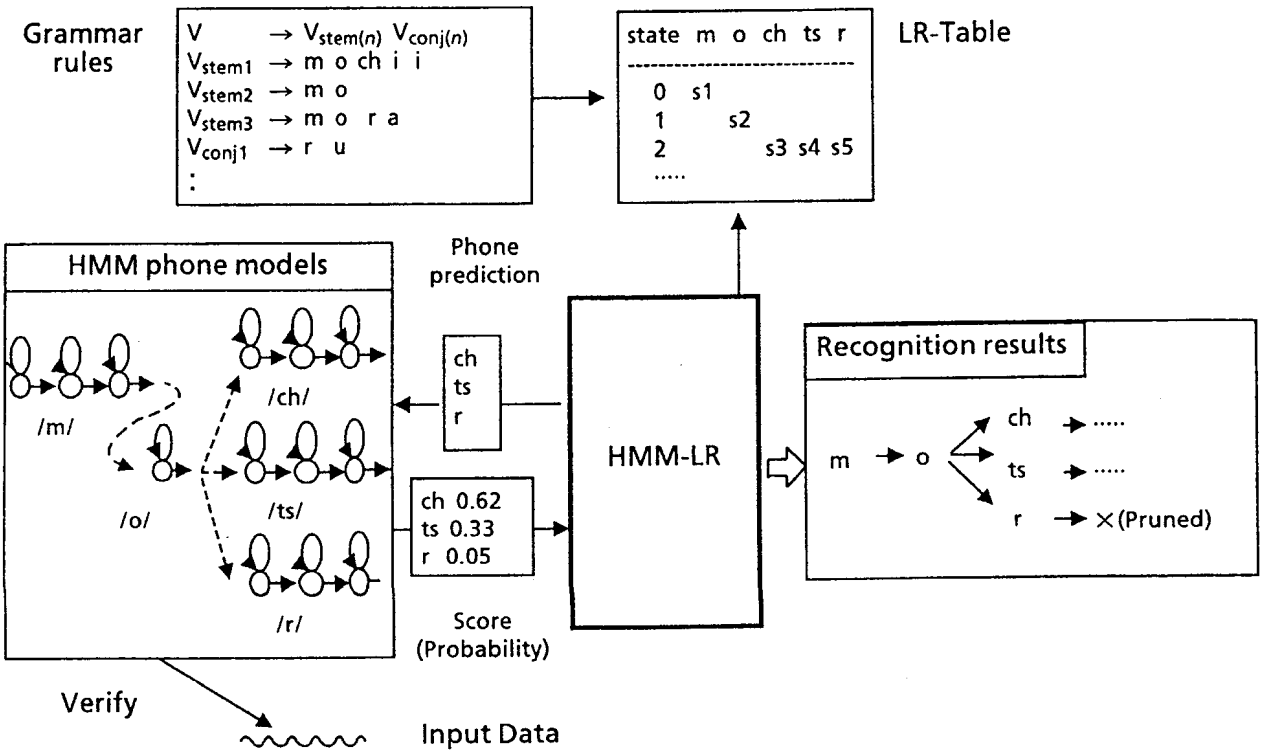


Figure 33.1

than independent phrases. Then, all outputs from the recognizer are almost syntactically correct.

There still, however, remain several ambiguities in the outputs from the recognizer because only syntactic constraints are used in the process. To solve this problem, not only the best candidate (the best hypothesis) but several candidates (n-best hypotheses) are output from the recognizer. The next step (Japanese analyzer) accepts such n-best hypotheses, and chooses the most plausible one that satisfies more accurate linguistic (syntactic and semantic) or even pragmatic constraints.

Spoken Language Translation

The style of spoken sentences is, especially in Japanese, quite different from that of written sentences. Spoken sentences include various intentional expressions or ellipses. To treat such sentences, a new method called the "intention translation method" (Kurematsu et al. 1991) was developed (figure 33.2).

An input utterance is analyzed by the analyzer based on the HPSG (and its Japanese version JPSG) grammar formalism and unification operation. In each lexical entry, syntactic, semantic and even pragmatic constraints are defined in the form of feature structures. In this paradigm, the inefficiency caused by the unification operation is the biggest issue, and various efforts have been made such as introducing medium-grained CFG rules (Nagata 1992) or implementing a quasi-destructive graph unification algorithm (Tomabechi 1992) to solve this issue. With these efforts, the processing time has been drastically decreased.

The next transfer component is composed of three phases: zero-anaphora resolution, illocutionary force type determination, and conversion of source-language semantics to target-language semantics.

In spoken Japanese, words that are easily inferable from the context tend to be omitted. In particular, "I" and "you" are seldom uttered explicitly. In many cases, such zero-anaphora can be resolved by the use of pragmatic information such as honorifics appearing in the sentence.

The semantics of an input utterance can be divided into two parts: an intentional content part and a propositional content part. Roughly speaking, the former part indicates the speaker's intention or attitude, and the latter is a neutral proposition. From the former, the illocutionary force type of the utterance is determined. Typical illocutionary force types are

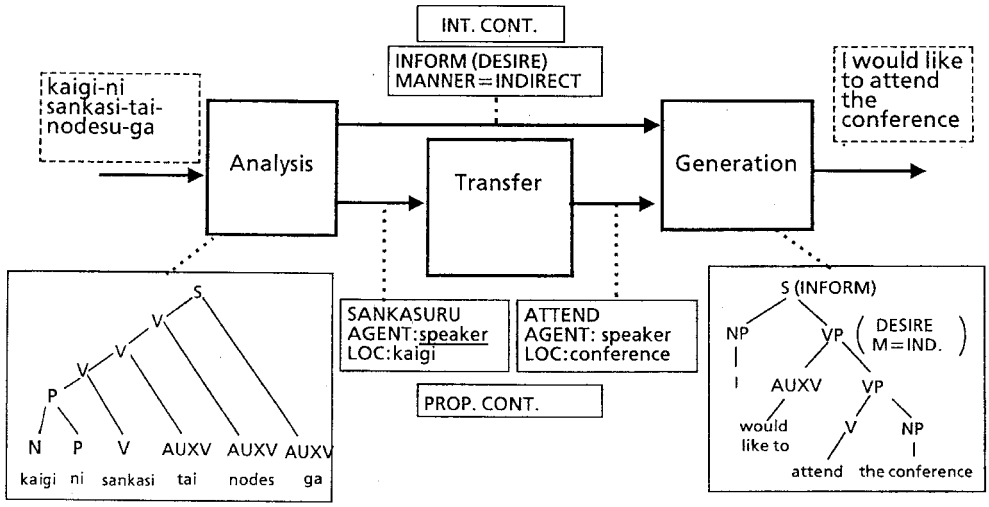


Figure 33.2

Table 33.1
Typical Illocutionary Force Type

Type	Explanation
PHATIC	Phatic expression such as those to open or close dialogue (*Hello*, *Thank you*)
INFORM	Inform a hearer of some facts
REQUEST	Request a hearer to carry out some action (*Please tell me* . . .)
QUESTIONIF	Yes/No question
QUESTIONREF	WH question

Table 33.2
Translation Examples

No.	Input	Translated Output
1	kaiginonaiyōnitsuiteoshi-etekudasai	Please tell me about the content of the conference
2	konkainokaiginowadaiwa tsūyakudenwadesu	The topic of the conference this time is interpreting telephony
3	watashiwaeigogazenzen-wakaranai nodesuga	I don't understand English at all
4	nihongoenodōjitsūyaku-woyōi shiteorimasu	Simultaneous interpretation into Japanese is available

shown in table 33.1. In other words, an intentional content part is converted to language-independent concepts. A propositional content part described in Japanese concepts is converted to corresponding target-language concepts.

The final component is generation. It accepts semantic feature structures in which both an illocutionary force type and a propositional content are described. The task of the generation system is to generate a syntax tree corresponding to input semantic feature structures. For this purpose, a set of subtrees annotated with semantic feature structures (called "phrase definition" or PD) is defined in the system. Such a PD is defined for each basic phrase structure as well as for each typical idiomatic expression of the target language. During generation, a set of PDs that can subsume the whole semantic feature structure of the input is selected, and combined by a unification operation. Finally, a succession of lexical words appearing at the bottom of the generated syntax tree is output as a result. Some translation examples are shown in table 33.2.

Speech Synthesis

In conventional speech synthesis, uniform speech units such as CVC (consonant-verbal-consonant) or VCV are prepared and the target speech is generated by connecting such units. In this approach, synthesized speech is not clear or natural enough because of distortion caused by the concatenation.

To improve the quality, a new method called Nyutalk has been developed (Sagisaka et al. 1992). In this method, various non-uniform units are extracted from a huge speech database and stored in a synthesis speech file.

For a sentence to be synthesized, the system dynamically selects the best combination (i.e. the one which makes the least distortion) from these non-uniform units. Finally, prosody of the output speech

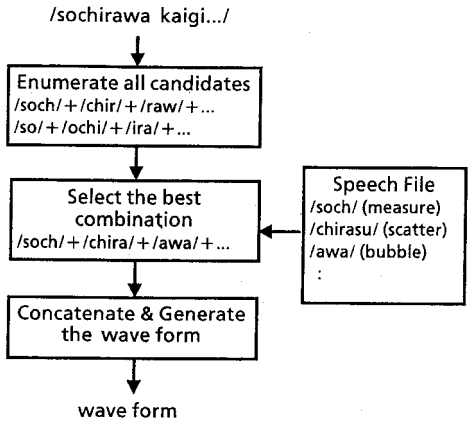


Figure 33.3

is controlled according to the syntactic structure of the sentence (figure 33.3).

ATR Speech Translation System

In 1989, ATR developed the first version of an experimental speech translation system from Japanese to English called SL-TRANS (Morimoto et al. 1992). Then, many improvements both in the mechanisms and the efficiency were made, and the final version called ASURA was implemented. Most of the technologies mentioned above are integrated in it except for an English speech synthesizer; for this component, a commercial English speech synthesizer (DecTalk) is used.

The target domain is inquiry about an international conference. Experiments have been conducted over 12 dialogues, which cover various varieties of topics such as inquiry on how to register, how to cancel, sightseeing tour, hotel arrangement, etc.

Two versions which differ in vocabulary size have been developed. The first one (hereafter called "the standard version") covers all of the 12 dialogues and about 700 words are defined in the system. The second one, whose lexicon size is about 1500, is more extended (we call it the "extended version"), and covers not only the dialogues, but also more than 90 percent of standard expressions in Japanese spoken sentences.

The overall speech-to-speech translation accuracy of these two versions is shown in table 33.3. With the standard version, more than 90 percent of the utterances are recognized and translated properly. With the extended version, the accuracy drops to about 63 percent. This is mainly due to an increase of ambiguity generated by the translation part. The processing

Table 33.3
Performance of ASURA

	Speech Recognition Rate	Translation Rate (System Total)
Standard Version	86.7% (1st) 92.5% (≥3rd)	90.3%
Extended Version	82.2% (1st) 90.7% (≥3rd)	63.3%

times of the standard version and the extended version are about 25 seconds and 50 seconds respectively (when two HP9000/750 are used). You might think that it is slightly too long. In the future, however, hardware innovations can still be expected and then near real-time processing might be achieved.

International Joint Experiment

ATR in Japan, CMU in the United States and Siemens Corporation/Karlsruhe University (KU) in Germany agreed to collaborate mutually in the area of speech translation, and started a consortium called C-STAR (Consortium for Speech Translation Advanced Research) a couple of years ago. The three parties decided to carry out an international joint experiment on an automatic interpreting telephone system, by interconnecting their speech translation systems. The parties shared equal responsibility; each site developed a speech recognition part and a speech synthesis part for its own language and a language translation part to the other two languages. In ASURA, all kinds of linguistic knowledge and the processing programs that use them are completely separated. Only transfer rules from Japanese to German and generation rules for German have newly been developed for Japanese–German translation. Other components such as Japanese analysis were used in common with those for the Japanese–English translation. Consequently, a Japanese–German translation system was developed in a very short time. The total system configuration for the experiment is shown in figure 33.4.

The experiment was conducted on January 28, 1993. Several dialogues out of 12 were used in the experiment. In addition to the speech translation system, a teleconference system was used so that the speaker could see what was going on at the other end. A large audience including the press and TV attended. As a whole, the experiment was successful and received a favorable evaluation.

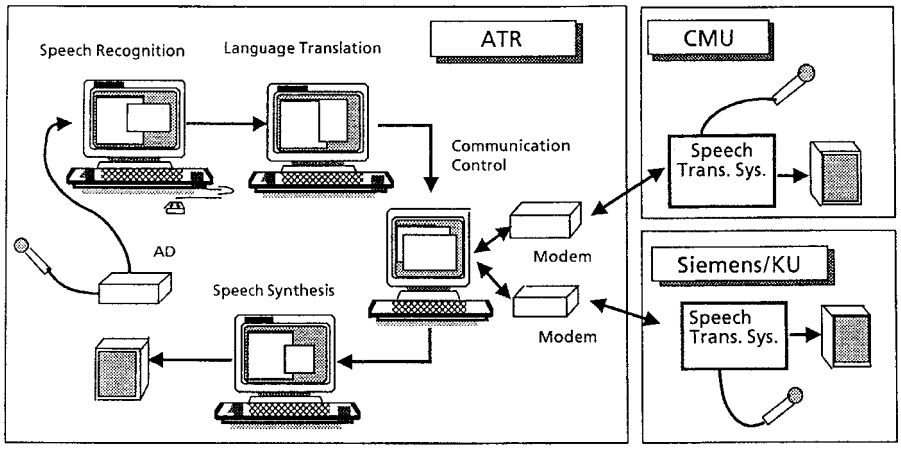


Figure 33.4

Further Enhancement and Extension

Interest in speech translation research has been growing; some work has been stimulated by the ATR interpreting telephony project. In the United States, several institutes such as CMU or Bell Labs have been making efforts in speech translation research. In Germany, the VERBMOBIL project was recently launched, whose aim is to develop a portable face-to-face speech translation system. The same kind of big national project has also started in South Korea.

Most of their goals are very exciting and ambitious, i.e. speech translation of spontaneous utterances.

At ATR, a new research organization (ATR Interpreting Telecommunications Research Laboratories) has recently been established supported by the Japan Key Technology Center and various private enterprises. It is the successor of the preceding project and will engage in basic research on advanced speech translation. In this section, a brief introduction of the new project will be given.

Objectives

The objective of the project is to develop key technologies for translation of spontaneously or naturally spoken utterances. Such utterances include wide varieties of speech and language phenomena, which have not so much been investigated until now. In speech, phenomena such as strong coarticulation, phone variation depending on an individual person, collapsed or missing phones, etc., will appear quite often. On the other hand, prosody plays an important role for conveying extra-linguistic information like a speaker's intention. As language phenomena, fragmental and strongly context dependent utterances, inversions, repeating or re-phrasing, ungrammatical expressions etc., will appear. The target area of the new project is shown in figure 33.5: the area covered area by the previous project is also indicated. In the following section, problems and approaches to be pursued are described.

Problems and Approaches

Recognition of Spontaneous Speech Speech recognition must be robust enough against both acoustic and linguistic variations. In acoustic research, much effort will be paid to developing more precise and robust allophonic models to cover wide acoustic variations. In that way, effects from the linguistic environment might be carefully considered. The recognition of a non-specific user's speech is also important. Some dynamic speaker adaptation mechanism must be established so as to eliminate the undesirable necessity of uttering a few words in advance. Adding to these themes, the problem of how to define and manage a language model should be investigated. Unnecessary or unimportant words such as *Uh* or *Oh* would be inserted frequently in spontaneous utterances. Difficulties may be the treatment of colloquial expressions like inversion and so on. Some management mechanism of a language model will be necessary, which will interact with a higher level language processing component and restructure the model dynamically according to progress of the dialogue.

Prosody Extraction and Control in Speech Synthesis Prosody such as intonation, power and speed will play an important roll in spontaneous speech. It helps not only to resolve ambiguities in sentence meanings, but sometimes to give the extra-linguistic

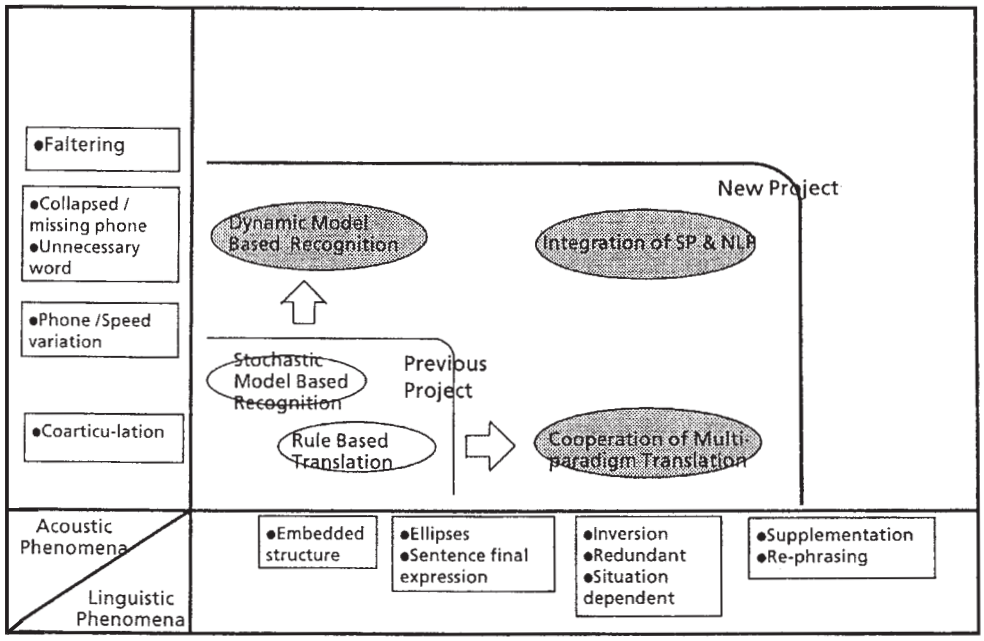


Figure 33.5

information such as the speaker's attitude, intention, or even emotion. Efforts will be made to establish an algorithm to extract prosody from speech and control it in speech synthesis.

Translation of Colloquial or Spontaneous Utterances
Most conventional translation is carried out by the use of several kinds of linguistic rules such as analysis, transfer or generation rules. It is based on the idea that all linguistic phenomena can be captured and written down as rules. However, we frequently observe various exceptional phenomena. However, a new translation paradigm, called the "example based translation" approach, has recently attracted considerable attention. It translates an input by using a set of translation examples each of which is very similar to a portion of the input. Such examples are extracted from a large bilingual corpus. This approach seems to be very promising for translation of spontaneous utterances. However, if such examples are used without any linguistic knowledge or principles, the results would be disappointing. We believe that the best way is to somewhat integrate a rule-based approach and an example-based approach; should these two algorithms collaborate with each other, the most likely translation is generated. At the same time, the dialogue situation at that point should be taken into consideration to translate the very context dependent expressions properly.

Integrated Control of Speech and Language Processing
Especially in spontaneous speech translation, the integrated control of speech and language processing becomes very important. Appropriate information necessary for language models should be provided to speech recognition from the language processing side, and speech information such as prosody should be provided to language processing from the speech processing side as well. At the same time, the status of the dialogue should be recognized and maintained properly. Such situational information would be about the environment (such as the domain or the subject of the dialogue), the participants' status (such as their intentional or mental states) and the dialogue progression status (such as the topic or the focus). Such information would be referred to by both the speech processing and the language processing. The overall image of the future system would be like figure 33.6.

Time Schedule and Managemental Issues
The term of the project is seven years (from March 1993 to March 2000) and the total budget is expected to be 16 billion yen, which is nearly the same amount as the previous project's budget. The number of researchers will be about 50.

Considering the importance of international cooperation, the project ardently wants to have good col-

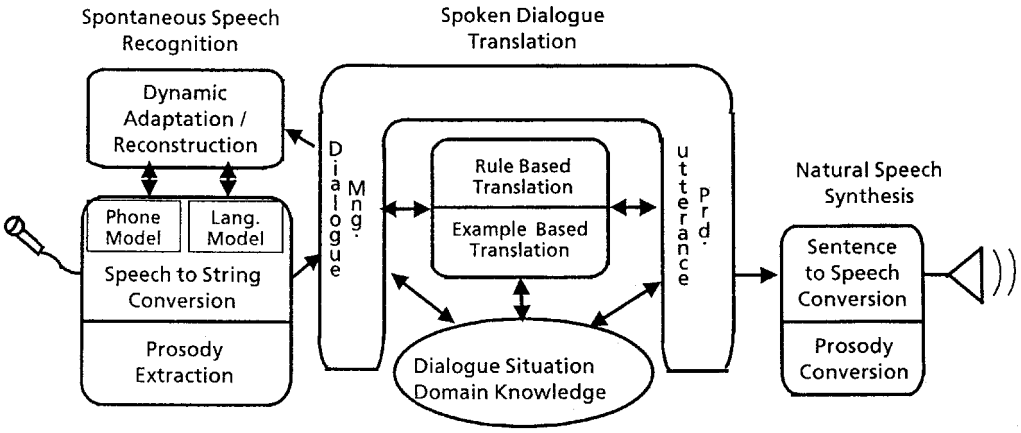


Figure 33.6

laborative relationships with outside active research organizations.

Conclusion

The main results achieved in the previous project are summarized below.

1. Componential technologies necessary for developing a speech translation system have intensively been studied and a prototype system has also been developed.

2. An international joint experiment, connecting ATR's, CMU's and Siemens/KU's systems, has been conducted and was successful.

3. Those efforts have shown the technical possibility of developing an "Interpreting Telephony System" in the near future.

The new project (following the previous project) was introduced. The mission of the project is to enhance and to extend the results attained in the previous project. We believe that these efforts will bring fruitful results, and make it possible for people in the world be able to speak freely without worrying about language differences at the beginning of the next century.

References

Kita, K., T. Kawabata, and H. Saito. 1989. HMM Continuous Speech Recognition Using Predictive LR Parsing. *Proceedings of ICASSP-89, IEEE International Conference on Acoustic, Speech and Signal Processing*.

Kurematsu, A., H. Lida, T. Morimoto, and K. Shikano. 1991. Language Processing in Connection with Speech Translation at ATR Interpreting Telephony Research Laboratories. *Speech Communication*, 10, 1–9.

Morimoto, T., M. Suzuki, T. Takezawa, G. Kikui, M. Nagata, and M. Tomokiyo. 1992. A Spoken Language Translation System: SL-TRANS2. *Proceedings of the Fifteenth International Conference on Computational Linguistics, COLING-92* (Nantes, France), 1048–1052.

Nagata, M. 1992. An Empirical Study on Rule Granularity and Unification Interleaving toward an Efficient Unification-Based Parsing System. *Proceedings of the Fifteenth International Conference on Computational Linguistics, COLING-92* (Nantes, France), 177–183.

Ohkura, K., M. Sugiyama, and S. Sagayama. 1992. Speaker Adaptation Based on Transfer Vector Field Smoothing with Continuous Mixture Density HMMS. *International Conference on Spoken Language Processing* (Banff, Canada).

Sagisaka, Y., N. Kaiki, N. Iwahashi, and K. Mimura. 1992. ATR NYU-Talk Speech Synthesis System. *International Conference on Spoken Language Processing* (Banff, Canada).

Saito, H., and M. Tomita. 1988. Parsing Noisy Sentences. *COLING Budapest: Proceedings of the 12th International Conference on Computational Linguistics*, 561–566.

Steer, M. G. 1987. A Speech Driven Language Translation System. *European Conference on Speech Technology* (Edinburgh).

Takami, J., and S. Sagayama. 1992. Successive State Splitting Algorithm for Efficient Allophone Modeling. *Proceedings of ICASSP-92. IEEE International Conference on Acoustic, Speech and Signal Processing*.

Tomabechi, H. 1992. Quasi-Destructive Graph Unification with Structure Sharing. *Proceedings of the Fifteenth International Conference on Computational Linguistics, COLING-92* (Nantes, France), 440–446.

This page intentionally left blank

34
The Stanford Machine Translation Project

Yorick Wilks

This paper describes a system of semantic analysis and generation, programmed in LISP 1.5 and designed to pass from paragraph-length input in English to French via an interlingual representation. A wide class of English input forms is covered, with a vocabulary initially restricted to a few hundred words. The distinguishing features of the translation system are:

It translates phrase by phrase, with facilities for reordering phrases and establishing essential semantic connectivities between them. These constitute the interlingual representation to be translated. This matching is done without the explicit use of a conventional syntax analysis.

The French output strings are generated without the explicit use of a generative grammar. This is done by means of stereotypes: strings of French words, and functions evaluating to French words, which are attached to English word senses in the dictionary and built into the interlingual representation by the analysis routines.

Introduction

The ongoing project to be described here aims to translate from English to French, using a reasonably wide vocabulary and paragraph-length texts, and at a later stage to "understand" the translated material, in the sense of being able to answer questions about it in an on-line context. The method to be used is a non-standard semantic analysis that has been applied to English texts of some length and complexity [13, 15].

It is the semantic approach that is intended to answer the question: "Why start Machine Translation (MT) again at all?" The generally negative surveys produced after the demise of most of the MT research of the fifties in no way established that a new approach was foredoomed to failure. At this time, it is easy to be unfair to the memory of that early MT work and to exaggerate the simplicity of its assumptions about language. But the fact remains that almost all of it was done on the basis of naive syntactic analysis and without the use of any of the developments in semantic structuring and description that have been noteworthy features of recent linguistic advances.

At this point a word of warning is appropriate about the semantic method used here. This is intended to be a practical talk, concerned with describing what is being done in a particular system, not with arguing abstractly for the advantages of systems based on conceptual connections over other contemporary but better-known approaches: this has been done elsewhere by writers such as Simmons [12], Quillian [9], Klein [3], Schank [11], as well as myself. I am not concerned, therefore, with arguing for a general method, nor shall I set out much in the way of the now familiar graph structures linking the items of example sentences in order to display their "real structure. I am concerned more with displaying the information structure I use, and how the system applies to certain linguistic examples to get them into the prescribed form for translation. The display of conceptual or dependency connections between items of real text will only be made in cases where unnecessary obscurity or complexity would be introduced by displaying the same connections between items of the interlingual representation.

This project is intended to produce a working artifact, not to settle general questions. However, because the territory has been gone over so heavily in the past years and because the questions still at issue seem to cause the adoption of very definite points of view, it is necessary to make certain remarks before beginning. In particular, different views are held at the present time on the question of whether the intermediate representation between two languages for MT should be logical or linguistic in form.

What the words in the last sentence, "logical" and "linguistic," actually mean is not as clear as might appear; for example, they are almost certainly not mutually exclusive; any "logical coding" of text will require a good deal of what is best called linguistic analysis in order to get the text into the required logical form: this could include coping with sense

ambiguity, clause dependency, and so on. On the other hand, few linguistically oriented people would deny the need for some analysis of the logical relations present in the discourse to be analyzed. However, for the purposes of the present project certain assumptions may safely be made:

1. Whatever linguists and philosophers may say to the contrary, it has never been shown that there are linguistic forms whose meaning cannot be represented in some logical system. Linguists often produce kinds of inferences properly made but not catered for in conventional existing calculi: for example, the "and so" inference in *I felt tired and went home*; but nothing follows to the effect that such an inference could not be coped with by means of a simple and appropriate adjustment in rules of inference.

2. Whatever logicians may believe to the contrary, it has never been shown that human beings perform logical transformations when they translate sentences from one language to another, nor has it ever been shown that it is necessary to do so in order to translate mechanically. To take a trivial example, if one wants to translate the English *is*, then for an adequate logical translation one will almost certainly want to know whether the particular use of *is* in question is best rendered into logic by identity, set membership, or set inclusion; yet for the purposes of translating an English sentence containing *is* into a closely related language like French, it is highly unlikely that one would ever want to make any such distinction for the purpose immediately at hand.

The above assumptions in no way close off discussion of the questions outstanding: they merely allow constructive work to proceed. In particular, discussion should be continued on: (a) exactly what the linguist is trying to say when he says that there are linguistic forms and common sense inferences beyond the scope of any logic, and (b) exactly what the logician is trying to say when he holds in a strong form the thesis that logical form is the basis of brain coding, or is the appropriate basis for computing over natural language.

On this subject we note the present conjunction of hitherto separate work: the extended set logic of Montague [7] that he claims copes with linguistic structure better than does MIT linguistics, and the work of G. Lakoff [4] which claims that the transformationalists in general and Chomsky in particular were always seeking for some quite conventional notion of logical form. However, these problems have not affected the development of our system which is designed to translate from one natural language to another and is potentially capable of question answering and the additional "understanding" that implies.

The coexistence of the two forms of coding, logical and linguistic, within a single system might preclude a way of testing the logicist and linguistic hypotheses about MT against each other. Such a test would be precluded because any translation into logic within such a system would have much of the work done by linguistic analysis; so there could be no real comparison of the two paths.

ENGLISH → PREDICATE CALCULUS REPRESENTATION → FRENCH
ENGLISH → LINGUISTIC CONCEPTUALIZATION → FRENCH

However, it might be possible to get translated output by each of the two paths in a single system and so give some rein to the notion of experimental comparison; I discuss this below.

The Structure of the Translation and Organization System

The diagram in figure 34.1 represents the overall structure of the system under construction.

I assume in what follows that processes 2, 4, and 5 are the relatively easy tasks—in that they involve throwing away information—while 1 and 3 are the harder tasks in that they involve making information explicit with the aid of dictionaries and rules.

With all the parts to the diagram and the facilities they imply (including not only translation of small texts via a semantic representation but also the translation of axioms in the predicate calculus (PC) into both natural languages) it is clear that input to the system must be rather restricted. However, there clearly are ways of restricting input that would destroy the point of the whole activity; for example, if we restricted ourselves to the translation of isolated sentences rather than going for the translation of paragraph-length texts. Whatever Bar-Hillel says to the contrary about MT being essentially concerned with utterances [1], I am assuming that the only sort of MT of interest here will be the translation of text.

The general strategy of translation is to segment the text in some acceptable way, produce a semantic representation as directly as possible, and generate an output French form from it. This involves mapping what I call semantic templates directly onto the clauses and phrases of English, and trying to map

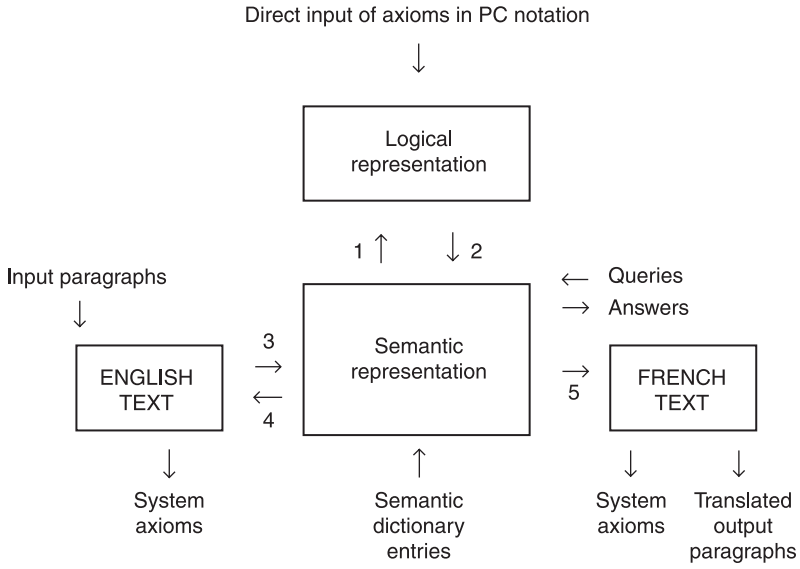


Figure 34.1

directly from the templates into French clauses and phrases, though with their relative order changed where necessary. I also assume that no strong syntax analysis is necessary for this purpose and that all that is necessary can be done with a good semantic representation—which leaves us with the question: what is in the semantic box, and how is it different from what is in the logic box?

I am using "semantic representation" narrowly to mean whatever degree of representation is necessary for MT—not necessarily for question-answering (that's what the logic box is for) or for theories of how the brain works. For this we may well not need the refinements of *is* mentioned earlier, nor, say, existential quantification or the analysis of presuppositions given by translation of definite descriptions. My main assumption here about the difference between the two boxes, logical and linguistic, is that an "adequate" logical translation makes all such matters explicit, and that is why it is so much more difficult to translate into the top box than the bottom one. But the difference between the two remains a pragmatic one, intended to correspond to two "levels of understanding" in the human being.

The Processing of English Text

The aim of the text-processing sections of the overall program is to derive from an English text an interlingual representation that has adequate expressivity as a representation from which: (1) output in another natural language can be computed, and (2) it can serve as an analysandum of predicate calculus statements about some particular universe.

The first pass made of the English input text is the fragmentation and reordering procedure, whose function is to partition and repack texts of some length and sentential complexity into the form most suitable for matching with the template forms mentioned above. This stage is necessary because, like all proposed coding schemes, logical, linguistic, or whatever, the template format is a more-or-less rigid one and the variety of natural language must be made to fit, if the system is to analyze anything more than simple example sentences.

The principal item of semantic structure used to analyze and express input text is the template. Templates are semantic frames, intended to express the messages or "gists" of the sentences and parts of sentences used in normal discourse. The system has an inventory of these templates available to it and seeks to match them with the fragments of the input text.

The template is of the basic form, subject-verb-object—or in semantic terms, actor-act-object—such as MAN HAVE THING, to be interpreted as "some human being possesses some object," and which would be matched as the bare template name of any sentence such as *John owns a car.* MAN, HAVE, and THING are interlingual elements, and MAN, for example, would be expected to be the principal, or head,

element for any semantic formula representing the English word *John* in the dictionary. Similarly, HAVE would be the head element in the appropriate semantic formula for *owns*, and so on. A simple matching algorithm would then be able to match the acceptable sequence of head elements from the template, MAN HAVE THING, onto a sequence of formulas drawn from the dictionary for the words of *John owns a car.*

The details of the matching algorithm are not a matter of concern here; what is important to see is that an algorithm for matching a bare three-element template onto a piece of language by inspecting just the head elements of formulas and searching for acceptable sequences of them will, in the course of making the match, select not only the head element of the word formula, but with it the whole formula of which it was the head, where "whole formula" is to be understood at this point as a coded form that expresses the whole content of the word sense in question. In the present case *John*, being a mere name, has no sense other than that it refers to a human being, and its whole formula would be simply (THIS MAN), which says no more than that.

One of the hypotheses at work here is that there is a finite inventory of templates adequate for the analysis of ordinary language—a usable list of the messages that people want to convey with ordinary language—and that in selecting those sequences of formulas for a fragment that are also template sequences (as regards their head elements) we pick up the formulas corresponding to the appropriate senses of the words of the fragment. This description is highly general; the details of the application of this method of analysis to complicated text appear in [15].

Moreover, it is assumed that any fragment of natural language can be named by (that is to say, matched with) at least one such bare template, and that the name will serve as a basic core of meaning for the purpose of translating the fragment. In other words, we translate from the complex interlingual representation of which the bare template MAN HAVE THING is the name simply because we know how to express as an algorithm the message "a person has a thing" in French. The template is thus an item, or unit, of meaning to be translated.

An example might help to give the general idea of what ties are established between text items by the matching routines described. Suppose we apply the above template to the sentence: *My brother owns a large car.* Let us suppose, furthermore, that we are not concerned with the problem of selecting the correct sense formulas, one corresponding to each word as it is used in the sentence. We shall make the simplifying assumption that each of those six words has only one sense entry in the dictionary, and that we are considering the relationships set up indirectly among the words by matching an interlingual representation onto the sentence.

From the point of view of the matching routine, the initial representation of the sentence is a string of six semantic formulas, whose details I shall discuss later. What matters at the moment is that the formula for *brother* has the head element MAN, just as did the one for *John*, and so on for *owns* and *car.* The formulas for *my* and *large* have the conventional head element KIND, since they specify what kind of thing is in question. The template-matching routine scans the formula string from left to right and is able to match the bare template MAN HAVE THING from the template inventory onto the formulas for *brother*, *owns*, and *car*, respectively, since those elements, in that order, are the heads of those formulas. Those three words are, as it were, the points in the sentence at which the template puts its three feet down.

So far, at the word level, ties that can be written as follows have been established:

brother ↔ owns ↔ car

These are much the same sort of ties that would be established at the word level by any system of conceptual semantic analysis applied to that sentence [11]. However, given that all realistically coded words in the dictionary would have many sense formulas attached to them, only certain selections of formulas would admit of being matched by an item in the template inventory. For example, in the sentence *This green bicycle is a winner*, the semantic formula for *winner* that has MAN as its head and means "one who wins" is never picked up by the matching routine simply because there is no bare template THING BE MAN in the inventory.

To return to the sentence *My brother owns a large car*, having matched on the bare template, the system looks at the three formulas it has tied together by means of their heads to see if it can extend the representation, top-down, by attaching other formulas and so create a fuller representation. In this case it looks from the formula for *brother* to the one that preceded it, the formula for *my.* This, it sees, can indeed qualify the formula for *brother*, and so it opens a list of formulas that can be tied onto this *brother* formula.

Repeating this process, we end up with an interlingual representation for the sentence in the following schematic form (which I shall call a full template, though we shall see later that the tied items are not simply formulas):

F [brother] ↔ F [owns] ↔ F [car]
↑ ↑
(F [my]) (F [large])

where both the horizontal and vertical directions represent dependency ties of the sort I have described, and F[x] stands simply for the interlingual formula for the English word x. Thus, the upwards vertical dependency is that of a list of qualifying formulas (empty in the case of *owns*) on the main formula.

The corresponding ties between the text words themselves established by this method are:

brother ↔ owns ↔ car ↔ a
↑ ↑
my large

A point that cannot have escaped the reader is that by having a rigid actor-action-object format for templates, one ignores the fact that many fragments of natural language are not of this form, regardless of how the initial input text is partitioned. This is indeed the case, but by using the notion of dummy parts of templates one can in fact put any text construction into this very general format. Since the analysis has no conventional syntactic base, the standard examples of syntactic homonymity, such as the various interpretations that can be thought up for *they are eating apples*, are represented only as differing message interpretations. So for that sentence we would expect to match at least the bare templates MAN DO THING and THING BE THING.

Fragmentation and Isolation

The fragmentation routine partitions input sentences at punctuation marks and at the occurrence of any of an extensive list of key words. This list contains almost all subjunctions, conjunctions, and prepositions. Thus the sentence *John is in the house* would be returned by such a routine as two fragments *(John is)* and *(in the house)*. With the first fragment the system would match MAN BE DTHIS, where the D of DTHIS indicates that, having failed to find any predicate after *is*, the system has supplied a dummy THIS to produce the canonical form of template.

If there is more than one available template to choose from, the preference is to the representation with the most conceptual connections (which can be thought of simply as the number of →s in the word diagrams) and the minimum number of dummies. For the fragment *in the house*, the matching routine finds itself confronted with a string of formulas, starting with one for *in*, that has PDO as its head. Prepositions are, in general, assimilated to actions and so have the P in the PDO of their heads to distinguish them from straightforward action formulas. In this case the matching routine inserts a dummy THIS as the leftmost member of the bare template, since it first encounters an action formula—headed by a PDO—as it scans the formula string from left to right, and *in the house* is finally matched with the bare template DTHIS PDO POINT. Thus the sentence *John is in the house* is partitioned into two fragments and matched with a semantic representation consisting of a string of two templates whose bare template names are MAN BE DTHIS and DTHIS PDO POINT.

Another example of fragmenting and matching is presented by what might conventionally be called noun phrases. If, after fragmenting, the system is presented with *The old black man* as a single fragment, it can supply two such dummies during the match and end up with a representation named by the bare template MAN DBE DTHIS.

The semantic connectivities described so far have been between formulas that correspond to words occurring in the same fragment of text. But not all semantic ties in a complex sentence will be internal to fragments—many will be between items occurring in different, and maybe not even textually contiguous, fragments. At a later point I shall discuss TIE routines whose function is to provide, in the full interlingual representation, those inter-fragment dependencies necessary for translation.

However, the major simplifying role of the fragmentation must not be lost in all this; it allows a complex sentence to be represented by a linear sequence of templates with ties between them, rather than by a far more complex hierarchical representation as is usual in linguistics.

The fragmentation, then, is done on the basis of the superficial punctuation of the input text and a finite list of keywords and keyword sequences, whose occurrence produces a text partition. Difficult but important cases of two kinds must then be considered. First, those where a text string is NOT fragmented even though a key word is encountered. Two intuitively obvious cases are non-subordinating uses of *that*, as in *I like that wine*, and prepositions functioning as "post verbs" as in *He gave up his post.* In these

cases there would be no fragmentation before the key words. In other cases text strings are fragmented even though a key word is NOT present. Four cases are worth mentioning:

1. *I want him to go* is fragmented as *(I want) (him to go)*. A boundary is inserted after any form of the words *say* and *want*, and a further boundary is inhibited before the following *to.* This seems intuitively acceptable, since *want* in fact subjoins the whole of what follows it in the sentence. We shall expect to match onto these fragments bare templates of the form MAN WANT DTHIS and MAN MOVE DTHIS respectively, where the first dummy THIS stands for the whole of the next template. The fragmentation functions operate at the lowest possible level of analysis; they inspect the semantic formulas given for a word in the dictionary, but they cannot assume that the choice among the formulas has been made.

A verb like *advise*, on the other hand, is not of this sort, since we can interpret *I advise him* in a way in which we cannot interpret *I want him* in the earlier case. So we would expect *I advise him to go* to receive no special treatment and to be fragmented as *(I advise him) (to go)*, on a key word basis.

2. Relative clauses beginning with *that* or *which* are located and isolated and then inserted back into the string of fragments at a new point. For example, *The girl that I like left* is fragmented as *(The girl left) (that I like PD)*, where the final period, PD, of the sentence is also moved to close off the sentence at a new point. Thus, the partition after *like* is made in the absence of any key word.

3. *The old man in the corner left* is, naturally enough, fragmented as *(The old man) (in the corner) (left)*. The breach made here between the actor and act of the sentence is replaced later by a tie (see below).

4. The sentences *John likes eating fish*, *John likes eating*, and *John began eating fish*, are all fragmented before *eating*, so that these forms are all assimilated to *John likes to eat fish* (which is synonymous with the first sentence above), rather than to *John is eating fish*, which would not be fragmented at all. In template terms *John is eating fish* is to be thought of as MAN DO THING, while *John likes fish* is MAN FEEL DTHIS+DTHIS DO THING, where the first DTHIS refers to the whole of the next template, and the second DTHIS stands in place of MAN (i.e., John).

Of is a key word that receives rather special treatment, and is not used to make a partition when it introduces a possessive noun phrase. After fragmentation, each fragment is passed through an ISOLATE function, which looks within each fragment and seeks for the right-hand boundaries of *of* phrases and marks them off by inserting a character "FO" into the text. Thus, *He has a book of mine* would be returned from the ISOLATE function as *He has a book of mine* FO. This is done in all cases except those like *I don't want to speak of him*, where *of* effectively functions as a post verb.

It may seem obvious enough why *of* phrases should remain within the fragment, since *of John* functions as does *John's*, but the demarcation of the phrase with the "FO" character can only be explained by considering the PICKUP and EXTEND routines.

PICKUP and EXTEND

The PICKUP routines match bare templates onto the string of formulas for a text fragment. As the routines move through the string of formulas, those contained between an OF and a FO are ignored for the purpose of the initial match. This ensures that "*of* phrases" are only treated as qualifiers. Thus, in the sentence *The father of my friend* FO *is called Jack*, the match would never try to make the head of the formula for *friend* into the root of a template matching the sentence, since it is sealed between an "*of*-FO" pair. To illustrate the results of applying PICKUP, I shall set down the bare templates that would be expected to match onto Nida and Taber's [8] suggested seven basic forms of the English indicative sentence. (In this talk I describe only the indicative mood as it is implemented in the trial version of this system. Queries and imperatives, like passives, are dealt with by the appropriate manipulation of the template order.)

In each case I give the basic sentence, the bare template, and a diagramatic representation of the corresponding dependencies implied between the text items, where "↔" again links those words on which the bare template is rooted or based, and "→" links a dependent word to its governor.

(1) John ran quickly — MAN MOVE DTHIS — John ↔ ran ↔ [DTHIS]
quickly → ran

(2) John hit Bill — MAN DO MAN — John ↔ hit ↔ Bill

(3) John gave Bill a ball — John ↔ gave ↔ ball
MAN GIVE THING — ↑ ↑
(to) Bill a

(The establishment of this dependency by EXTEND is discussed below.)

(4) John is in the house.
MAN BE DTHIS DTHIS PBE THING
John ↔ is ↔ [DTHIS] [DTHIS] ↔ in ↔ house
↑
the

(5) John is sick — John ↔ is ↔ sick
MAN BE KIND

(6) John is a boy — John ↔ is ↔ boy
MAN BE MAN — ↑
a

(7) John is my father — John ↔ is ↔ father
MAN BE MAN — ↑
my

A natural question at this point is, what exactly is this inventory of bare templates to be used in the analysis of input languages? No detailed defence is offered of the inventory used, nor, I believe, can one be given. The fact is that one uses the inventory that seems empirically right, revises it when necessary, and concludes that, alas, is how things must be in the real world of practical language analysis.

The inventory used can be reconstructed from the table of rules set out below in Backus-Naur Form. It is set out in terms of the action designating semantic elements, such as FORCE, and the classes of substantive designating elements (such as *SOFT meaning STUFF, WHOLE, PART, GRAN, and SPREAD) that can precede such an action as a subject, and follow it as an object to create a three-element bare template.

```
<bare template>::= <*PO><DO><*EN>|
<*PO><CAUSE><*EN>|<*PO><CHANGE><*EN>|
<*AN><FEEL><*MA>|<*EN><HAVE><*EN>|
<*AL><PLEASE><*AN>|<*AL><PAIR><*EN>|
<*PO><SENSE><*EN>|<*PO><WANT><*EN>|
<*PO><USE><*EN>|<*PO><TELL><*MA>|
<*PO><DROP><*EN>|<*PO><FORCE><*EN>|
<*EN><MOVE><DTHIS>|<*PO><GIVE><*EN>|
<*AL><WRAP><*EN>|<*AN><THINK><*AM>|
<*SO><FLOW><DTHIS>|<*PO><PICK><*EN>|
<*PO><MAKE><*EN>|<*AL><BE><same member of *AL
  as last occurrence>
<*AL>::=<DTHIS|THIS|MAN|FOLK|GRAIN|PART|WORLD|
  STUFF|THING|BEAST|PLANT|SPREAD|LINE|ACT|
  STATE>
  (*AL means all substantive elements)
<*EN>::=<DTHIS|THIS|MAN|FOLK|GRAIN|PART|STUFF|
  THING|BEAST|PLANT|SPREAD|LINE>
  (*EN means elements that are entities)
<*AN>::=<MAN|FOLK|BEAST|GRAIN>
  (*AN means animate entities, GRAIN is used as
  the main element for social organizations,
  like the Red Cross)
<*PO>::=<DTHIS|THIS|MAN|FOLK|GRAIN|PART|STUFF|
  THING|ACT|BEAST|PLANT|STATE>
  (*PO means potent elements, those that can
  designate actors. The class cannot be
  restricted to *AN since rain wets the grass
  and the wind opens doors)
<*SO>::=<STUFF|PART|GRAIN|SPREAD>
<*MA>::=<ACT|SIGN|STATE>
  (*MA designates mark elements, those that can
  designate items that themselves designate
  like thoughts and writings)
```

I have distorted BNF very slightly to write the bare templates containing BE in a convenient and perspicuous form. The forms containing MOVE and FLOW also contain a DTHIS (i.e., they are "dummy templates") indicating that there cannot be objects in those bare templates. Thus, MOVE is used only in the coding of intransitive actions and not to deal with such sentences as *I moved all the furniture round the room.*

There are dummy templates not included in this list—several occur in the description of the Nida and Taber sentences above. The remaining rules specifying them are intuitively obvious, but may be found in detail in [15], with important ancillary rules which specify when dummies are to be generated in matching sentences. Naturally, a dummy MAN BE DTHIS is generated for the first fragment of *(John is) (in the house)* simply because a proper three-element bare template cannot be fitted onto the information available. But in other cases, where a three-element template can be fitted, dummies are generated as well, since subsequent routines to be described may prefer the dummy to the bare template. For example, in the analysis of the first fragment of *(The old transport system) (which I loved) (in my youth) (has been found uneconomic)* a reasonably full dictionary will contain formulas for the substantive sense of *old* and the action sense of *transport.* Thus, the actor-action-object template FOLK CAUSE GRAIN can be fitted on here but will be incorrect. The dummy GRAIN

DBE DTHIS will also be fitted on and will be preferred by the EXTEND procedures described below. Such slight complexity of the basic template notion is necessary if so simple a concept is to deal with the realities of language. This matter is described is greater detail in [15].

The matching by PICKUP will still, in general, leave a number of bare templates attached to a text fragment. It is the EXTEND routines, working out from the three points at which the bare template attaches to the fragment, that try to create the densest dependency network possible for the fragment, and thus reduce the number of templates matching a fragment.

In order to show more clearly how EXTEND does this, it is necessary to say more about the semantic formulas which make up the full template. A semantic formula expresses the meaning of one sense of a natural language word in the dictionary. It is made up of left and right parentheses and of semantic elements. The latter include THING, STUFF, MAN, etc., for basic items in the world; FORCE, CAUSE, DROP, CHANGE to describe basic kinds of actions; and so on. The formulas are binarily bracketed pairs of whatever depth of nesting is necessary to express the meaning of a particular word sense. The formulas are made up, and interpreted, with a dependency of the left element, or bracket group, upon the corresponding right-hand element or bracket group in every case.

So, (MAN KIND) would be interpreted as "of a human sort;" it is a formula for "human" used as a qualifier. In ((MAN FEEL) CAUSE) the dependency within the inner bracket is of an actor-act type, whereas that within the outer bracket—of (MAN DO) on CAUSE—is of the object-of-action on act type. So the whole sub-formula is to be interpreted as "causes a person to feel something," and we would therefore expect to find this sub-formula within any formula for, say, *torment.* (There are restrictions on the ways in which the elements can combine contained in a table of "scope notes" for the system of coding: for example, CAUSE cannot be anything but an action, so ((MAN DROP) CAUSE) could not be the specification of a sort of cause, but only the causing of something. The most important element in a formula is its rightmost one, or head, with which PICKUP connects formulas for words to templates for whole fragments.)

Formulas that can qualify any other substantive formula have the head KIND, and those that can qualify actions have the head HOW. Most action formulas have as head DO, BE, MOVE (*run,* for example), or GIVE. GIVE verbs are important in that they can function in the representation of action constructions like *He left John his watch*, where an indirect object of an action can appear without any preceding preposition. GIVE verbs function in much the same way as TRANS verbs in Schank's analysis [11], and the appearance of GIVE as a formula for, say, the action *left* primes the system to expect such an indirect object. The verb *tell* also has GIVE as the head of its principal formula, since it can participate in such indirect object constructions as *John tells me a story.* The lack of necessary connection between the English word *tell* and the interlingual element TELL is brought out by the fact that the formula head of *tell* is not TELL but GIVE. In the case of *say*, on the other hand, the head of its main formula is TELL, since it cannot occur in the GIVE-type constructions.

Most substantive formulas have as their heads such elements as MAN, STUFF, THING, ACT (for abstract substantives which are the result of action, such as *adjustment*); STATE (abstract substantives such as *friendship*, *happiness*), GRAIN (abstract substantives any sort of structure such as *system*) and so on. A formula for a substantive is assumed to be singular unless the element MUCH is its first item at the TO level.

Action formulas can specify a preferred class of actors, or of objects of the action, or both. Preferred actors are specified by FOR and preferred objects by TO. So then the formula for the action *talk* will contain the pair (MAN FOR), since most things that talk are human, and if there is a possibility of setting up a dependency with a human actor, the system will take it. The restriction cannot be absolute in this, or most other cases, since machines and dogs talk, in fable if not in fact. The important facility is to be able to PREFER the usual, if a representation for it is available, but to be able to accept the unusual if necessary.

The syntax of the action formula is as follows: (X FOR) or (X TO) appears as the first item at the top level of the action formula if appropriate—in LISP terminology the pair is simply CONS'd onto the verb formula. If both are appropriate, as in a formula for *interrogate*, then the (X TO), for the objects, is CONS'd first, and appears at one level lower in the nesting of the formula than the (X FOR), specifying the preferred actors. Thus the formula for *interrogate* would read: ((MAN FOR) ((MAN TO) (TELL FORCE))). The preferred substantives, or classes of

them, for qualifiers are indicated in an extension of this notation, by including (X FOR) as the first item at the top level in the formula for a qualifier.

In order to keep a small usable set of interlingual semantic elements, and to avoid arbitrary extensions of the list of elements, many notions are coded by conventional sub-formulas: (FLOW STUFF) is used to designate liquids, for example, and (WHERE SPREAD) to code spatial area of any sort.

The role of EXTEND was discussed in general terms above: it inspects the strings of formulas that replace a fragment, and seeks to set up dependencies of formulas on each other. It keeps a score as it does so, and in the end selects the structuring of formulas with the most dependencies, on the assumption that it is the right one (or ones, if two or more structurings of formulas have the same dependency score).

The dependencies that can be set up are of two sorts: (1) those between formulas whose heads are part of the bare template, and (2) those of formulas whose heads are not in the bare template upon those formulas whose heads are in the bare template.

Consider the sentence *John talked quickly*, for which the bare template would be MAN TELL DTHIS, thus establishing, at the word level, the dependency *John talked* [DTHIS]. Now suppose we expand out from each of the elements constituting the bare template in turn. In the formula for *talked* there is the preference for an actor formula whose head is MAN—since talking is generally done by people.

This preference is satisfied here; we can think of it as establishing a word dependency of *John* on *talked*, which is a type (1) dependency. Expanding again from the element TELL, we have a formula for *quickly* whose head is HOW, and HOW-headed formulas are proper qualifiers for actions. Hence we have been able to set up the following diagrammatic dependency at the word level:

John ⇆ talked ↔ [DTHIS]
↑
quickly

(where "⇆" indicates a bare template connectivity strengthened by a direct semantic dependency—springing from the preference of *talked* for a human actor in this case), and we would score two for such a representation. Furthermore, the formulas having type (B) dependence would be tied in a list to the main formula on which they depend. The subtypes of dependence are as follows:

A. Among the formulas whose heads constitute the bare template

1. preferred subjects on actions *John talked*

2. preferred objects of actions on actions *interrogated a prisoner*

B. Of formulas not constituting bare templates on those that do

1. qualifiers of substantives on substantives *red door*

2. qualifiers of actions on actions *opened quickly*

3. articles on substantives *a book*

4. *of*-FO phrases on substantives *the house of my father* FO

5. qualifiers of actions on qualifiers of substantives *very much*

6. post verbs on actions *give up*

7. indirect objects on actions *gave John a . . .*

8. auxiliaries on actions *was going*

9. *to* on infinitive form of action *to relax*

The searches for type (B) dependencies are all directed in the formula string in an intuitively obvious manner: 1, 3, 4, 5, and 8 go leftwards only; 6 and 7 go rightwards only; and 2 goes rightwards and leftwards.

The purpose of the score of dependencies established will become clear if we consider an example of (B) (7): the indirect object construction. Let us take the sentence *John gave Mary the book*, onto which the matching routine PICKUP will have matched two bare templates as follows, since it has no reason to prefer one to the other:

John gave Mary the book
MAN GIVE MAN
MAN GIVE THING

EXTEND now seeks for dependencies, and since the formula for *gave* has no preferred actors or objects, the top bare template cannot be extended at all, and so scores zero. In the case of the lower bare template, then, a GIVE action can be expanded by any substantive formula to its immediate right which is not already part of the bare template. Again, *book* is qualified by an article, which fact is not noticed by the top bare template. So then, by EXTENDing we have established in the second case the following dependencies at the word level and scored two (of the "→" dependencies):

John ↔ gave ↔ book
↑ ↑
Mary the

Thus the second representation is preferred. This is an application of the general rule referred to earlier as "pick the most connected representation from the fragment."

The auxiliary of an action also has its formula made dependent on that of the appropriate action and the fact scored, but the auxiliary formulas are not listed as dependent formulas either. They are picked up by EXTEND and examined to determine the tense of the action. They are then forgotten and an element indicating the tense is CONS'd onto the action formula. In its initial state the system will recognize only four tenses of complex actions:

PRES: does hide/is hiding/did hide/are hiding/am hiding
IMPE: was hiding/were hiding
PAST: did hide/had hidden
FUTU: will hide/will be hiding/shall hide/shall be hiding

In the case of the negative of any of these tenses the word *not* is forgotten, and an atom NPRES, NIMPE, NPAST, or NFUTU attached to the appropriate action formula instead. At present the system does not deal with passives, though I indicate later how they are dealt with within the template format.

Even when the representation with the densest dependency has been found, there may still be more than one representation with that score for a given fragment. So, in the case of *The man lost his leg* there may well be two representations of this sentence with the same dependency score, one corresponding to each of two different senses of *leg*—one as a part of a body, and one as an inanimate thing that supports some other thing (as in *piano leg*). There is a further routine in EXTEND, called into play in such cases, that attempts to establish additional "semantic overlap" of content both between the actor and object formulas of the template, and between each of the three main formulas of the template and its qualifiers. If any can be found, the additional dependencies are used to choose among representations that have achieved the same score in the EXTEND routines described earlier. So, in the present case, the formula for "leg of a person" would be expected to contain the sub-formula (MAN PART), whereas the formula for "piano leg" would not, and this connectivity with the initial formula of the template, whose head was MAN, would suffice for one representation to be chosen in preference to the other, again on the principle of preferring the most connected representation.

The third and last pass of the text applies the TIE routines, which establish dependencies between the representations of different fragments. Each text fragment has been tied by the routines described so far to one or more full templates, each consisting of three main formulas to each of which a list of dependent formulas may be tied. The interlingual representation consists, for each text fragment, of one full template together with up to four additional items of information called key, mark, case, and phase respectively. The interlingual representation also contains the English name of the fragment itself.

The **key** is simply the first word of the fragment, if it occurs on the list of key words; or, in the cases of *that* and *which* a key use of the word.

The **mark** for a given key is the text word to which the key word ties the whole fragment of which it is the key. So, in (He came home) (from the war), the mark of the second fragment is *came* and the second fragment is tied in a relation of dependence to that mark by the key *from.* Every key has a corresponding mark, found by TIE, unless (a) the key is *and* or *but* or (b) the fragment introduced by the key is itself a complete sentence, not dependent on anything outside itself. The notion will become clearer from examining the example paragraph set out below.

From the point of view of the present system of analysis, the case of a fragment, if any, generally expresses the role of that fragment in relation to its key and mark: it specifies the sort of dependence the fragment has upon its mark. In general, case markers are attached to fragments on the basis of the key and the mark. It may be that no case is finally assigned to a fragment, though it will be if a fragment is introduced by a preposition. The cases are, in a sense, a cross classification of prepositions, whose correct rendering into, say, French is so vital for adequate translation.

The provisional working list of cases and the English prepositions that can introduce them is as follows:

RECEIVER: to, from, for
INSTRUMENTAL: with, by
DIRECTION: to, from, towards, outof, for
POSSESSION: with
LOCATION (space and time): at, by, near, after, in, during, before

CONTAINMENT: in
SOURCE: outof, from
GOAL: to, at
OBJECT (as in (I want) (her to leave)): no key word necessary

The case analysis routines in TIE work by considering the above classification of prepositions in reverse, as it were: thus, in *(He struck the boy) (with a stick)*, TIE locates the *with* and finds in the stereotypes for *with* that *with* can introduce either a POSSESSIVE or INSTRUMENTAL fragment. If, for example, an INSTRUMENTAL case is in question it will expect a preceding action whose head is DO, CAUSE, or FORCE, and will also expect a substantive in the fragment it introduces whose head is THING. In the case mentioned, it finds these conditions satisfied, since the head of the appropriate formula for *stick* is THING, and so it ties the second fragment to the mark *hit* and assigns the INSTRUMENTAL case to the second fragment as a description of that tie.

In any other situation, where these criteria are not satisfied, the fragment introduced by *with* is tied to the immediately preceding substantive, and the case POSSESSIVE is assigned to the tie, as in *(He struck the boy) (with long hair)*. In one special class of cases, the POSSESSIVE case is assigned even though a THING substantive is found in the "object position" of the second template following a DO, CAUSE or FORCE action in a preceding template. These are the cases where the object is a part of the substantive previously mentioned. For, even though a leg is a THING, we would want to assign a POSSESSIVE case to the second template of the pair *(He hit the boy) (with the wooden leg)*. How this TIE is obtained algorithmically is discussed in detail in the section after the description of STEREOTYPES.

This procedure can be thought of as an ambiguity resolution of the prepositions, which has not been dealt with at all by the PICKUP routines, since prepositions are inserted into the formula strings as a single formula and are never considered ambiguous at that stage. The TIE routines also resolve other semantic ambiguity not dealt with by the PICKUP routines. If our last example had been *(He struck the boy) (with a bar)* we would have expected there to be at least two formulas for *bar* still in play: corresponding to the heads THING and POINT—the latter corresponding to the place sense of *bar*. Hence, there would still be two full templates matching the latter fragment at this stage, both considered by TIE, which would refer the template containing the sense of *bar* coded with the head THING, since only in that case could a dependency tie be made (to *hit* in another fragment, in this case) on the basis of information extracted from the formulas.

Phase notation is merely a code to indicate in a very general way to the subsequent generation routines where in the "progress of the whole sentence" one is at a given fragment. A phase number is attached to each fragment on the following basis by TIE, where the stage referred to applies at the BEGINNING of the fragment to which the number attaches.

0: main subject not yet reached

1: subject reached but not main verb

2: main verb reached but not complement or object

3: complement or object reached or not expected

The Interlingual Representation

What follows is a version of the interlingual representation for a paragraph, designed to illustrate the four forms of information—key, mark, case, and phase. The schema below gives only the bare template form of the semantic information attached to each fragment—the semantic formulas and their pendant lists of formulas that make up the full template structure are all omitted.

(LATER CM) → (PLUS TARD VG)
nil:nil:nil:O:No Template

(DURING THE WAR CM) → (PENDANT LA GUERRE VG)
DURING:GAVEUP:location:0:DTHIS PBE ACT

(HITLER GAVE UP THE EVENING SHOWINGS CM) → (HITLER RENONCA AUX REPRESENTATIONS DU SOIR VG)
nil:nil:nil:O:MAN DROP ACT

(SAYING) → (DISANT)
nil:HITLER:nil:3:DTHIS DO DTHIS

(THAT HE WANTED) → (QU'IL VOULAIT)
THAT:SAYING:object:3:MAN WANT DTHIS

(TO RENOUNCE HIS FAVORITE ENTERTAINMENT) → (RENONCER A SA DISTRACTION FAVORITE)
TO:WANT:object:3:DTHIS DROP ACT

(OUTOF SYMPATHY) → (PAR SYMPATHIE)
OUTOF:RENOUNCE:source:3:DTHIS PDO SIGN

(FOR THE PRIVATIONS OF THE SOLDIERS PD) → (POUR LES PRIVATIONS DES SOLDATS PT)
FOR:SYMPATHY:recipient:3:DTHIS PBE CT

(INSTEAD RECORDS WERE PLAYED PD) → (A LA PLACE ON PASSA DES DISQUES PT)
INSTEAD:nil:nil:O:.MAN USE THING (comment:template active)

(BUT) → (MAIS)
BUT:nil:nil:O:No Template

(ALTHOUGH THE RECORD COLLECTION WAS EXCELLENT CM) → (BIEN QUE LA COLLECTION DE DISQUES FUT EXCELLENTE VG)
ALTHOUGH:PREFERRED:nil:O:GRAIN BE KIND

(HITLER ALWAYS PREFERRED THE SAME MUSIC PD) → (HITLER PREFERAIT TOUJOURS LA MEME MUSIQUE PT)
nil:nil:nil:O:MAN WANT GRAIN

(NEITHER BAROQUE) → (NI LA MUSIQUE BAROQUE)
NEITHER:MUSIC:qualifier:O:DTHIS DBE KIND

(NOR CLASSICAL MUSIC CM) → (NI CLASSIQUE VG)
NOR:INTERESTED:nil:O:GRAIN DBE DTHIS

(NEITHER CHAMBER MUSIC) → (NI LA MUSIQUE DE CHAMBRE)
NEITHER:INTERESTED:nil:0:GRAIN DBE DTHIS

(NOR SYMPHONIES CM) → (NI LES SYMPHONIES VG)
NOR:INTERESTED:nil:O:GRAIN DBE DTHIS

(INTERESTED HIM PD) → (NE L'INTERESSAIENT PT)
nil:nil:nil:l:DTHIS CHANGE MAN

(BEFORE LONG THE ORDER OF THE RECORDS BECAME VIRTUALLY FIXED PD) →
(BIENTOT L'ORDRE DES DISQUES DEVINT VIRTUELLEMENT FIXE PT)
BEFORELONG:nil:nil:O:GRAIN BE KIND

(FIRST HE WANTED A FEW BRAVURA SELECTIONS) → (D'ABORD IL VOULAIT QUELQUES SELECTIONS DE BRAVOURE)
nil:nil:nil:O:MAN WANT PART

(FROM WAGNERIAN OPERAS CM) → (D'OPERAS WAGNERIENS VG)
FROM:SELECTIONS:source:3:DTHIS PDO GRAIN

(TO BE FOLLOWED PROMPTLY) → (QUE DEVAIENT ETRE SUIVIES RAPIDEMENT)
TO:OPERAS:nil:3:MAN DO DTHIS (comment: shift to active template again may give a different but not incorrect translation)

(WITH OPERETTAS PD) → (PAR DES OPERETTAS PT)
WITH:FOLLOWED:nil:3:DTHIS PBE GRAIN

(THAT REMAINED THE PATTERN PD) → (CELA DEVINT LA REGLE PT)
nil:nil:nil:O:THAT BE GRAIN (comment: no mark because *that* ties to a whole sentence)

(HITLER MADE A POINT OF TRYING) → (HITLER SE FAISAIT UNE REGLE D'ESSAYER)
nil:nil:nil:O:MAN DO DTHIS (comment: some idiom recognition essential to cope with this)

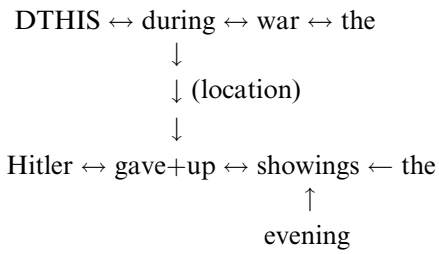
(TO GUESS THE NAMES OF THE SOPRANOS) → (DE DEVINER LES NOMS DES SOPRANOS)
TO:TRYING:object:2:DTHIS DO SIGN

(AND WAS PLEASED) → (ET ETAIT CONTENT)
AND:HITLER:nil:3:DTHIS BE KIND

(WHEN HE GUESSED RIGHT CM) → (QUAND IL DEVINAIT JUSTE VG)
WHEN:PLEASED:location:3:MAN DO DTHIS

(AS HE FREQUENTLY DID PD) → (COMME IL LE FAISAIT FREQUEMENT PT)
AS:GUESSED:manner:3:MAN DO DTHIS

It is assumed that those fragments that have no template attached to them, such as (LATER), can be translated adequately word-for-word. Were it not for the difficulty involved in reading it, we could lay out the above text so as to display the dependencies implied by the assignment of cases and marks at the word level. These would all be of dependencies of whole fragments on particular words. For example, the relation of just the first two fragments appears as:



This intermediate stage is an arbitrary one in the English–French processing that is useful to examine at the surface level. It is often supposed that an intermediate stage like the present interlingual representation must contain "all possible semantic information" in some explicit form if it is to be adequate. But the quoted words are not, and cannot be, well-defined with respect to any coding scheme. What is the case is that the interlingual representation must contain sufficient information to admit of the formal manipulations, adequate for producing translations in natural or formal languages. The IR need not contain any particular explicit information about a text. The real restriction is that in its creating no information should have been thrown away that will later turn out to be important; one of the difficulties of English–French MT is the need to EXTEND and make explicit in the French things that are not so in the English.

Consider the sentence *The house I live in is collapsing*, which contains no subjunction *that*, though in French it must be expressed explicitly, as by *dans laquelle.* There need not be any representation of *that* anywhere in the IR. All that is necessary is that the subordination of the second fragment to the mark *house* be coded, and generation procedures which know that in such cases of subordination an appropriate subjunction must occur in the French output. It is the need for such procedures that constitutes the sometimes awkward expansion of English into French, but the need for them in no way dictates the explicit content of the IR.

The Dictionary Format

The dictionary is essentially a list of pairs of semantic formulas (each corresponding to one sense of an English word), and of explanations of that sense. By "explanation" I mean not simply an English word or phrase, such as was used in earlier versions of this system of analysis [15], but what I shall call a French stereotype. For example, one sense of the English word *colorless* might have appeared in the dictionary as:

(((((WHERE SPREAD) (SENSE SIGN))
NOTHAVE) KIND)
(COLORLESS AS NOT HAVING THE
PROPERTY OF COLOR))

The first half of the pair, the formula, expresses the fact that being colorless means not having a spatial (WHERE SPREAD) sensory property (SENSE SIGN). The second half of the pair is a sense explanation in English that contains the name of the word and serves to distinguish that particular sense of *colorless* from other senses—such as one about human character.

But the senses of the English words may equally well be explained and distinguished by means of their French equivalents, at least in cases where the notion of "a French equivalent to an English word" is an appropriate one. So, for example, the French words *rouge* and *socialiste* might be said to distinguish two senses of the English word *red*, and we might code these two senses of *red* in the dictionary by means of the sense pairs:

(((WHERE SPREAD) KIND) (RED (ROUGE)))
((((WORLD CHANGE) WANT) MAN) (RED
(SOCIALISTE)))

The French words *rouge* and *socialiste* are enclosed in list parentheses because they need not have been, as in this case, single French words. They could be French word-strings of any length: for example, the qualifier sense of *hunting* as it occurs in a *a hunting gun* is rendered in French as *de chasse*; hence, we would expect (HUNTING (DE CHASSE)) as the right-hand member of one sense pair for *hunting.*

This simplified notion of stereotype is adequate for the representation of most qualifiers and substantives. Below I shall generalize to the notion of a **full stereotype** adequate for the representation of prepositions and actions, in which there may be more than one list after the English word name in the right-hand member of the sense pair. Moreover, they will be lists in which functions will occur as well as the names of French words.

We should pause at this point to see what the notions of **sense pair** and stereotype are doing for us in the system. Earlier I described the structure of a full template as made up of formulas and lists of formulas. But these would more accurately have been described as sense pairs and lists of sense pairs; the analysis routines, in fact, build into the template not just the formulas, but the **whole sense pairs**, of which

the formulas are the left-hand members. Hence, the full template already contains the French equivalents of the English words in the fragment. Moreover, the stereotypes for actions and prepositions contain not only French equivalents but implicit rules for assembling these equivalents to generate French output: the generation routines never need consult an English-French dictionary. The full template may appear to be a complex and cumbrous item of information, containing as it does not only a conceptual semantic representation of English text, but also French output forms and implicit generation rules; still, the avoidance of repeated consultation of a large dictionary of forms and rules in LISP format is no small compensation.

The full stereotype, then, may contain not only French words but also predicates and functions of interlingual items whose values are always French word strings, or a blank item, or NIL. The notion of "interlingual item" here covers not only the interlingual elements that make up the formulas, but also the names of the cases abbreviated to a standard four-letter format, for example; RECE, INST, DIRE, POSS, LOCA, CONT, SOUR, GOAL, OBJE, QUAL (see the list of cases given earlier).

The general form of the stereotype is a list of predicates, followed by a string of French words and functions that evaluate to French words, or to NIL (in which case the stereotype fails). The functions may also evaluate to blank symbols for reasons to be described.

The predicates, which occur only in preposition stereotypes, normally refer, respectively, to the case of the fragment containing the word and to its mark. If both these predicates are satisfied, the program continues on through the stereotype to the French output.

Let us consider the verb *advise*, rendered in its most straightforward sense by the French word *conseiller.* It is likely to be followed by two different constructions, as in the English: (1) *I advise John to have patience*, and (2) *I advise patience*.

Verb stereotypes contain no predicates, so we might expect the most usual sense pair for *advise* to contain a formula followed by

```
(ADVISE (CONSEILLER A (FN1 FOLK MAN))
        (CONSEILLER (FN2 ACT STATE
        STUFF)))
```

The role of the stereotypes should by now be becoming clear: in generating from, in this case an action, the system looks down a list of stereotypes tied to the sense of the action in the full template. If any of the functions it now encounters evaluate to NIL, the whole stereotype containing the function fails and the next is tried. If the functions evaluate to French words, they are generated along with the French words that appear as their own names, like *conseiller.*

The stereotypes do more than simply avoid the explicit use of a conventional generative grammar; they also direct the production of the French translation by providing complex context-sensitive rules at the required point without any search of a large rule inventory. This method is, in principle, extendable to the production of reasonably complex implicit rephrasings and expansions, as in the derivation of *si intelligent soit-il* from the second fragment of *(No man) (however intelligent) (can survive death)*, given the appropriate stereotype for *however.*

Preposition stereotypes are, in general, more complex than those for actions, but before illustrating them I should mention a point that arises in connection with stereotypes and their relation to the enumeration of the senses of the input. As I have described the dictionary so far, many output stereotypes may be attached to one sense of an English word, that is to a single semantic formula. In the example sentences above, *advise* is taken as being used in the same sense in the two sentences, even though different constructions follow the word in the two cases. So the notion of stereotype in no way corresponds to that of word sense. Indeed, the notion of word-sense is extremely unclear and resistant to any formal analysis.

In the case of prepositions, I take them as having only a single sense each, even though that sense may give rise to a great number of stereotypes. Let us consider, by way of example, *outof* (considered as a single word) in the three sentences:

(1) (It was made) (outof wood)

(2) (He killed him) (outof hatred)

(3) (I live) (outof town)

It seems to me unhelpful to say that here are three senses of *outof*, even though its occurrence in these examples requires translation into French by *de*, *par*, and *en dehors de*, respectively, and other contexts would require *parmi* or *dans.* Given the convention for stereotypes described earlier for actions, let us set down stereotypes that would enable us to deal with these cases:

(S1) ((PRCASE SOUR) (PRMARK *DO) DE (FN1 STUFF THING))
(S2) ((PRCASE SOUR) (PRMARK *DO) PAR (FN2 FEEL))
(S3) ((PRCASE LOCA) EN DEHORS DE (FN1 POINT SPREAD))

Here *DO indicates a wide class of action formulas: any, in fact, whose heads are not PDO, DBE, or BE.

When the program enters the second fragment of *(It was made) (outof wood)* it knows from the whole interlingual representation described earlier that the case of that fragment is SOURCE and its mark is *made.* The mark word has DO as its head, and so the case and mark predicates PRCASE and PRMARK in the first stereotype are both satisfied. Thus, *de* is tentatively generated from the first stereotype and FN1 is applied, because of its definition to the object formula in this template, the one for *wood.* The arguments of FN1 are STUFF and THING, and the function finds STUFF as the head of the formula for *wood* in the full template, is satisfied, and generates *bois* from the stereotype for *wood.*

In the case of the second fragment of *(He killed him) (outof hatred)*, the two predicates of the first stereotype for *outof* would again be satisfied, but (FN1 THING STUFF) would fail with the formula for *hatred*, whose head is STATE. The next stereotype (S2) would be tried; the same two predicates would be satisfied, and now (FN2 FEEL) would be applied to (NOTPLEASE (FEEL SAME)), the formula for *hatred.* But FN2 by its definition does not examine formula heads, but rather seeks for the containment of one of its arguments within the formula. Here it finds FEEL within the formula and so generates the French word stereotype for *hatred.*

Similar considerations apply to the third example sentence involving the LOCATION case, though in that case there would be no need to work through the two SOURCE stereotypes already discussed, since, when a case is assigned to a fragment during analysis, the only stereotypes left in the interlingual representation are those that correspond to the assigned case. In the case of fragments with a key, TIE routines search the stereotypes for the key until they find one that matches the fragment and its mark except with respect to case. So in the sentence *(I live) (outof town)*, the analysis routines assign LOCATION to the second fragment in the first place, because they locate in the third stereotype for *outof* a formula for the object of the preposition whose head is POINT.

The Generation of French

Much of the heart of the French generation has been described in outline in the last section, since it was impossible to describe the dictionary and its stereotypes without describing the generative role of the stereotypes.

To complete this sketch we need some description of the way in which generations from the stereotype of a key and of the mark for the same fragment interlock—the mark being in a different fragment—as control flows backwards and forwards between the stereotypes of different words in search of a satisfactory French output. There is not space available here for description of the bottom level of the generation program—the concord and number routines—which in even the simplest cases needs access to mark information (e.g., in locating the gender of *heureux* in *(John seems) (to be happy)* translated as *Jean semble être heureux*).

Again, much of the detailed content of the generation is to be found in the functions evaluating to French words that I have arbitrarily named FN1, . . . , etc. Some of these seek detail down to gender markers. For example, one would expect to get the correct translations *Je voyageais en France* but . . . *au Canada* with the aid of functions, say, FNF and FNM that seek not only specific formula heads but genders as well. So, among the stereotypes for the English *in* we would expect to find (given that formulas for land areas have SPREAD as their heads): . . . A (FNM SPREAD)) and . . . EN (FNF SPREAD)).

It is not expected that there will be more than twenty or so of these inner stereotype functions in all, though it should be noted at this point that there is no level of generation that does not require quite complicated semantic information processing. I have in mind here what one might call the bottom level of generation, the addition and compression of articles. An MT program has to get *Je bois DU vin* for *I drink wine*, but *J'aime LE vin* for *I like wine.* Now there is no analog for this distinction in English and nothing about the meanings of *like* and *drink* that accounts for the difference in the French in a way intuitively acceptable to the English speaker. At present we are expecting to generate the difference by means of stereotypes that seek the notion USE in the semantic

codings which will be located in *drink* but not in *like,* and to use this to generate the *de* where appropriate.

The overall control function of the generation expects five different types of template names to occur:

(1) *THIS *DO *ANY where:
*THIS is any substantive head (not DTHIS)
*DO is any real action head (not BE, PDO, DBE)
and *ANY is any of *DO or KIND or DTHIS

With this type of template the number, person, and gender of the verb are deduced from the French stereotype for the subject part.

(2) type *THIS BE KIND is treated with type 1.

(3) DTHIS *DO *ANY

These templates arise when a subject has been split from its action by fragmentation. The mark of the fragment is then the subject. Or the template may represent an object action phrase, such as a simple infinitive with an implicit subject to be determined from the mark.

(4) *THIS DBE DTHIS

Templates of this type represent the subject, split off from its action, represented by a type (2) template, above. The translation is simply generated from the stereotype of the subject formula, since the rest is dummies, though there may arise cases of the form DTHIS DBE KIND where generation is only possible from a qualifier, as in the second fragment of *(I like tall CM) (blond CM) (and blue-eyed Germans).*

(5) DTHIS PDO *REAL

Templates of this type represent preposition phrases, and the translation is generated as described from the key stereotype, after which the translation for the template object is added (*REAL denotes any head in *THIS or is KIND).

The general strategy for the final stages of the program is to generate French word strings directly from the template structure assigned to a fragment of English text. The first move is to find out which of the five major types of template distinguished above is the one attached to the fragment under examination.

For a fragment as simple as *John already owns a big red car*, the program would notice that the fragment has no mark or key, hence, by default, the generation is to proceed from a stereotype which is a function of the general type of the template attaching to the fragment. The bare name of the template for this one fragment sentence is MAN HAVE THING, and inspection of the types above will show this to be a member of type (1). The stereotype is a function, let us say FTEMP, of that template type and, to conform with the general format for stereotypes described earlier, this can be thought of as being one of the stereotypes for the "null word."

In this case, in the generation of French, function FTEMP evaluates to a French word string whose order is that of the stereotypes of the English words of the fragment. This order is directed by the presence of the first type of template, comprising an elementary sequence subject-action-object. This is done recursively so that, along with the French words generated for those English words whose formulas constitute the bare template (i.e., *John*, *own*, and *car*), formulas are generated that are merely dependent on the main formulas of the template—in this case the formulas for *already*, *big*, and *red.*

If complex stereotypes are located while generating for any of the words of the fragment, then generation from these newly found stereotypes immediately takes precedence over further generation from the last stereotype at the level above! "Complex" simply means full stereotypes which have constituents that are functions as well as French words.

Now suppose we consider the two-fragment sentence *I order John to leave.* The fragments will be presented to the generation program in the form described earlier: with key, mark, case, and phase information attached to each fragment:

(I order John) nil:nil:nil:0
(to leave) to:order:OBJE:2

Also attached to the fragments will be full templates whose bare template names in this case will be MAN TELL MAN and DTHIS MOVE DTHIS, respectively.

The generation program enters the first fragment, which has no mark or key; so it starts to generate, as before, from a stereotype for the null word, which again is one for the first template type. This gets the subject right: *je* from the stereotype for *I*, later to be modified to *j'* by the concord routine. It then enters the stereotypes for the action, the first being (ORDONNER A (FN1 MAN FOLK)).

The head of the formula for *John* is MAN. FN1 here is an arbitrary name for a function that looks into the formula for the object place of a template and, if the head of that formula is any of the function

arguments, it returns the stereotype value of that formula. In this case FN1 is satisfied by *John*, thus that stereotype for *order* is satisfied. The program generates from it the sequence *ordonner à Jean*, giving the correct sequence *Je$ ordonner$ à Jean* (where $ indicates the need for further minor processing by the concord routine). The stereotype has now been exhausted—nothing in it remains unevaluated or ungenerated; similarly, the fragment is exhausted, since no words remain whose stereotypes have not been generated, either directly or via the stereotype for some other word, and so the program passes on to the second fragment.

The program enters the second fragment and finds that it has a mark, namely *order.* It then consults the stereotype in hand for *order* in the first fragment to see if it was exhausted. It was, and so the program turns to the stereotypes for *to*, the key of the second fragment. Among those whose first predicate has the argument OBJE will be the stereotype ((PRCASE OBJE) (PRMARK FORCE TELL) DE (FNINF *DO)).

The head of the current formula for *order*, the mark of the second fragment is FORCE, and PRMARK seeks and compares its arguments with the head of the mark formula. The predicates are seen to be satisfied and the program generates *de* after seeing that FNINF is satisfied, since an action formula for *leave* follows, whose head, MOVE, is in the class *DO.

FNINF on evaluation finds the implicit subject of the infinitive. That is unnecessary here, but would be essential in examples only slightly more complex, such as *Marie regrette de s'être rejouié trop tôt.* Finally, FNINF itself evaluates to the French stereotype selected for *leave.* This might give rise to more searching if the use of *leave* dictated its own sequents, as in *I order John to leave by the first train.* Here, however, the evaluation terminates immediately to *partir*, since the sentence stops. Thus the correct French string *Je$ ordonne$ à Jean de partir* has been generated.

The last example was little more than a more detailed redescription of the processes described in the dictionary section, in connection with the example *I advise John to have patience.* However, now that we have dealt fully with a fairly standard case and shown the recursive use of stereotypes in the generation of French on a fragment-by-fragment basis, we can discuss a final pair of examples in which a more powerful stereotype can dictate and take over the generation of other fragments.

If we were to consider in detail the generation of French for the two-fragment sentence *(I throw the ball) (outof the window)*, we should find the process almost identical to that used in the last example. In this case, too, the main stereotype used to generate the French for the first fragment is that of the action—*throw* in this case—and the stereotype for *throw* is exhausted by the first fragment, so that nothing in that stereotype causes the program to inspect the second fragment.

Now consider, in the same format, *(I drink wine) (outof a glass)*. Following the same procedures as before, we shall find ourselves processing the stereotype for *drink*, which reads (BOIRE (FN1 (FLOW STUFF)) (FNKl SOUR PDO THING) ↑ DANS (FNX2 THING) ↑, where "↑" indicates a **halt-point**. The program begins to generate tentatively, evaluating the functions left to right and being prepared to cancel the whole stereotype if any one of them fails. FN1 is applied to the formula for *wine* and specifies the inclusion in its formula, not of one of two elements, but of the whole conventional sub-formula for liquids (FLOW STUFF). This, it finds, is satisfied, and so evaluates to *vin*, to be modified by concord to *du vin.*

The program now encounters FNX1, a function which by definition applies to the full template for some following fragment. At this point the program evaluates FNX1 which returns a blank symbol if and only if it finds a following fragment with a SOURCE case and a template. The last two elements, of whose bare name are PDO THING, i.e., it is a preposition-type fragment with a physical object as object. This situation would not obtain if the sentence were *I drink the wine outof politeness.* If FNX1 is satisfied, as in this case, it causes the generation from this stereotype to halt after generating a blank symbol. Halting in an evaluation is to be taken as quite different from both exhausting (all functions evaluated to French word strings or a blank) and failing (at least one function evaluates to NIL).

The main control program now passes to the next fragment, in this case *outof a glass.* It asks first if it has a mark, which it has, namely *drink*, and looks at the stereotype in hand for the mark to see if it is exhausted, which it is not, merely halted. The program therefore continues to generate from the same stereotype, for *drink*, producing *du vin*, then *dans*, followed by the value of FNX2, namely *verre*, thus giving the correct translation *Je $bois$ du vin dans un verre.*

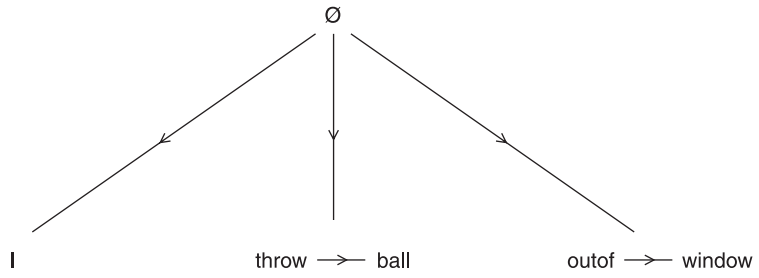


Figure 34.2

The important point here is that the stereotypes for the key to the second fragment, *outof*, are never consulted at all. The translations for all the words of the second fragment will have been entered via a stereotype for the previous fragment, the one for *drink*. The advantage of this method will be clear: because it would be very difficult, conceptually and within the framework described, to obtain the translation of *outof* as *dans* in this context from the stereotype for *outof*, since that translation is specific to the occurrence of certain French words, such as *boire*, rather than to the application of certain concepts. In this way the stereotypes can cope with linguistic idiosyncrasy as well as with conceptual regularity. It should be noted, too, that since *dans* is not generated until after the halted stereotype restarts, there is no requirement that the two example fragments be contiguous. The method I have described could cope just as well with *(I drink the wine) (I like most) (outof a silver goblet)*.

For clarification about what words are generated through the stereotypes for what other words, a diagram follows in which lines connect the English word through whose stereotype a generation is done to the word, for which output is generated. All generations conventionally start from ϕ, the null word mentioned above; it is, by convention, the word for which the five basic stereotypes are the stereotype. The more straightforward case *(I threw the ball) (outof the window)* would be generated as in figure 34.2.

Articles are omitted for simplicity. In this case the new fragment starting with *outof* returns to ϕ to begin generating again. In the more complex case *(I drink wine) (outof a glass)*, the generation pattern would be as in figure 34.3.

The general rule with action stereotypes, then, is that the more irregular the action, the more information goes into its stereotype and the less is needed in the stereotypes for its sequents. So, for example, there is no need for a stereotype for *outof* to contain

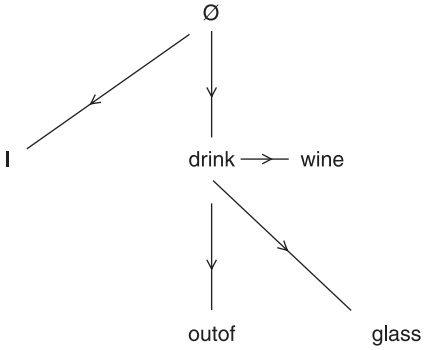


Figure 34.3

DANS at all. Again, just as the regular case *I order John to leave* produced the translation *J'ordonne à Jean de partir* by using the stereotype for the key *to*, the less regular case *I urge John to leave*, which requires the quite different construction *J'exhorte Jean à partir*, would be dealt with by a halting stereotype for *urge* whose form would be (EXHORTER (FN1 MAN FOLK) (FNX1 OBJE *DO) ↑A (FNXINF *DO)).

In this case, the stereotype for *to* would never be consulted at all.

Finally, it should be admitted that in the actual analysis and generation system, two items described, "case" and "mark," shrink in importance, though by no means disappear. Their role has been overstressed in this paper, in order to make a clear distinction between the analysis and generation routines and so present a clear interlingual representation whose format is independent of the algorithmic techniques employed. What I sought to avoid was any reference to a "seamless computational whole" all of whose levels seem to presuppose all of the other levels, and which even if it works, cannot be inspected or discussed in any way.

The assignment of the case and mark information demands access to the French stereotypes. It would

clearly be absurd to consult the stereotypes to assign this information and then, later, consult them again in order to make use of it in the generation of French. In fact, the analysis and generation routines fuse at this point, and the case and mark are located during the generation of the French output. The change in the format that this requires is that the mark predicate PRMARK is not now simply a predicate that checks whether the *already assigned* mark for the fragment in hand meets the specification: it is a predicate that at the same time actively seeks for a mark meeting that specification. And, as with the stereotype functions already described, the failure to find such a mark fails the whole stereotype containing it. There will now be a number of mark predicates fulfilling different roles. The case predicate, conversely, is not diversified but vestigial, because there is now no *previously assigned* case to a fragment for the predicate to check, and the case is now just a label in the dictionary of stereotypes to aid the reader.

A last, quick look at a previous example should make all this clear. Consider again *(He hit the boy) (with the wooden leg)* as contrasted with the alternative second fragments *(with a stick)* and *(with long hair)*. Let us consider the analysis routines terminating with the provision of full templates for fragments (and phase information), and let us consider everything that follows that a French generation.

Let us now consider the generation program entering the second fragment, armed with the following list of stereotypes for *with*:

((PRMKOB *ENT) (POSS) A (FN *ENT))
((PRMARK *DO) (INST) AVEC (FN THING))
(PRMARK *ENT) (POSS) A (FN *REAL))

PRMKOB is a directed predicate that seeks for a mark in a preceding fragment (within a range of two fragments). It looks only at candidates whose heads are in the class *ENT, that is, THING, MAN, FOLK, BEAST, or WORLD; entities that can in some sense have parts in the same sense the heads ACT, STATE, POINT, etc., are not attached to word senses that we can speak of as having parts. PRMKOB compares the formulas for potential marks in the third, object, template position of preceding fragments with the formula for the object in the template for the fragment in hand. And it is true if and only if the latter formula indicates that it ties to a word sense that can be a part of the entity tied to the "candidate mark" formula.

So, in the case of *(He hit the boy) (with the wooden leg)* PRMKOB finds itself comparing the formulas for *boy* (head MAN) and *leg* (which contains the subformula (MAN PART). In this case PRMKOB is satisfied and the generation continues through the first stereotype, correctly generating *à* for *with* and then the output for *wooden leg*. The *REAL in the function in the first stereotype merely indicates that any object in that fragment should then have its stereotype generated (any substantive head is in the class *REAL), because its appropriateness has already been established by the satisfaction of PRMKOB.

Following exactly the procedures described in other examples, it will be seen that *(with a stick)* fails the first but is translated by the second stereotype, while *(with long hair)* fails the first two but is correctly generated by the third.

References

[1] Bar-Hillel, Y. 1970. Some Reflections on the Present Outlook for High-Quality Machine Translation. Mimeo, University of Texas.

[2] Bierwisch, M. 1970. Semantics. In J. Lyons (ed.), *New Horizons in Linguistics*. London: Pelican.

[3] Klein, S., et al. 1968. The Autoling System, Tech. Report. #43, Computer Science Dept., University of Wisconsin.

[4] Lakoff, G. 1970. Linguistics and Natural Logic. *Studies in Generative Semantics #1*, University of Michigan, Ann Arbor.

[5] McCarthy, J., and Hayes, P. 1969. Some Philosophical Problems from the Standpoint of Artificial Intelligence. *Machine Intelligence 4*, Edinburgh.

[6] Michie, D. 1971. On Not Seeing Things. Experimental Programming Reports #22, University of Edinburgh.

[7] Montague, R. 1970. English as a Formal Language. In *Linguaggi nella Societa e nella Tecnica*, Milan.

[8] Nida, E., and Taber, C. 1969. *The Theory and Practice of Translation*. Leiden: Brill.

[9] Quillian, R. 1969. The Teachable Language Comprehender. *Communications of the ACM*.

[10] Sandewall, E. 1971. Representing Natural Language Information in Predicate Calculus. *Machine Intelligence 6*, Edinburgh.

[11] Schank, R. 1971. Finding the Conceptual Content and Intention of an Utterance in Natural Language Conversation. *Proceedings of the 2nd Joint International Conference on Artificial Intelligence*. London.

[12] Simmons, R. 1970. Some Semantic Structures for Representing English Meanings, Tech. Report #NL-1, University of Texas at Austin.

[13] Wilks, Y. 1968. On-Line Semantic Analysis of English Texts. *Machine Translation and Computational Linguistics*.

[14] ———. 1971. Decidability and Natural Language. *Mind*, 80, 497–516.

[15] ———. 1972. *Grammar, Meaning and the Machine Analysis of Natural Language*, London: Routledge.

[16] Winograd, T. 1971. Procedures as a Representation for Data in a Computer Program for Understanding Natural Language, Project MAC Memo MAC TR-84, Massachusetts Institute of Technology.

35
The Textual Knowledge Bank: Design, Construction, Applications

Victor Sadler

Introduction

The concept of the Textual Knowledge Bank (or TKB) is one which grew out of research into machine translation at the BSO software company in the Netherlands over the period 1985–1990. This was the time span of a state-subsidized R&D project leading to a prototype MT system under the name of DLT (Distributed Language Translation). An overview of the DLT project—now suspended pending fresh funding—can be found in Witkam (1988). The development of the Textual Knowledge Bank concept is described in Sadler (1989), and its first pilot implementation in Sadler and Vendelmans (1990).

Basically, the TKB concept is simple enough. It represents a way of storing full text, not as an extended string of characters, but as a grammatically and referentially coded tree structure in which the nodes are linguistic objects on various levels and from which the original character string can be reconstructed at any level (figure 35.1). The aim is to make the knowledge contained in ordinary texts accessible to the computer—without formalizing the linguistic knowledge into rules and without building an abstract knowledge representation divorced from the linguistic level. To this end, the text has to be structured: first by identifying its components (words, morphemes or whatever); second by drawing syntactic relations between those components (dependency parsing); and third by drawing reference relations between components which in one way or another refer to the same thing (anaphora, etc.). In this way, it was argued, both the linguistic and the non-linguistic knowledge (knowledge of the world) required for natural language processing could be combined into a single knowledge source.

Figure 35.2 illustrates the TKB structure, comprising both syntactic and referential links, for the following pair of sentences:

(1) Use the Delete option to delete individual documents. The owner and other sharers cannot access those documents.

Where bilingual or multilingual applications are concerned, a further dimension is added to the structure. For MT purposes, for instance, parallel texts (translations) in different languages are first structured in the way described above, and then additional, bilingual relations are drawn between equivalent units in the two parallel structures (figure 35.3). The result is termed a Bilingual Knowledge Bank (or BKB).

Figure 35.4 illustrates by means of an example sentence the coupling of two (monolingual) TKBs into a (bilingual) BKB. In this figure, the dependency structure of the English and French sentences is shown in a different graphical form from that of figure 35.2, with (sub)trees defined by (boxes within) boxes. It will be clear that a BKB can function as a kind of bilingual (and bidirectional) dictionary, with an abundance of contextual examples.

As far as linguistic knowledge is concerned, the motivation behind the construction of a large database such as a TKB is the conviction that rule-based systems have proved inadequate for NLP purposes, and that analogical or "example-based" or "memory-based" techniques are needed instead–or possibly in addition (the hybrid approach). This is a conviction which developed at BSO in the course of the DLT project and which seems to be echoed more and more frequently by other researchers. To quote just one advocate of this view (Skousen 1989:100):

> Speakers generally lack the ability to make explicit the rules that govern their behavior. An analogical approach suggests that the reason for this inability is not that the rules are somehow inaccessible, but instead that the rules don't actually exist.

Human beings are usually able to use language effectively without necessarily learning explicit grammatical rules, even where such rules can be established with reasonable confidence. So, surely computers too can be equipped with the ability to behave in a similar fashion. If so, their first requirement must be a large body of examples on which to base their

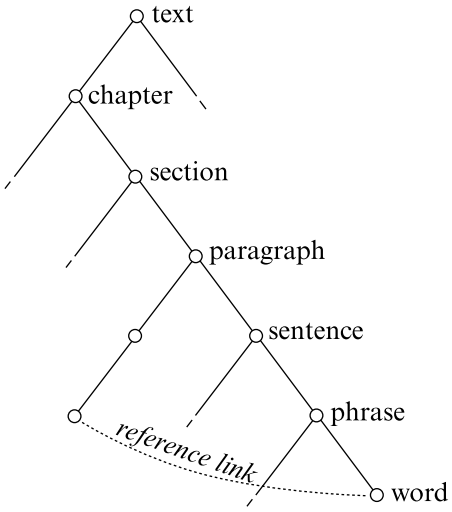


Figure 35.1
Conceptual view of a Textual Knowledge Bank.

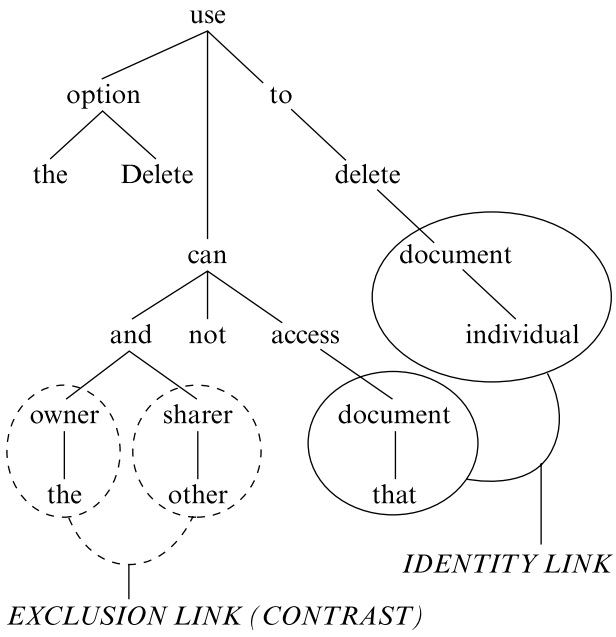


Figure 35.2
TKB structure for example (1).

analogical reasoning. And, given that most applications will require knowledge of discourse phenomena as well as sentence-level and word-level phenomena, the body of examples should consist not just of isolated phrases or sentences, but of continuous texts.

Of course, where naturally regular phenomena are concerned, it may be worth while, for the sake of efficiency, to extract the rules from the structured corpus off-line, and then at runtime first apply the rules obtained before resorting to analogy. This possibility, rather than arguing against the use of a corpus-based knowledge bank, actually provides a further motive for building one. It is not practicable to extract all the rules of a language from its corpus, even if the rules could be enumerated, because it is extremely difficult to decide in advance which rules, or what kind of rules, may be needed for a particular application. So the on-line corpus remains a valuable knowledge source. But where a language lends itself to description in terms of rules, there the corpus-based approach offers a means of objectively checking any proposed rules against actual observed usage, or alternatively of discovering rules automatically.

As for non-linguistic knowledge, there is general agreement that sophisticated NLP is not possible without some form of knowledge of the world. The question on which there is very little agreement is how to represent that knowledge. There is growing interest of late in using lexical relations derived from corpora (such as verb–subject and verb–object patterns) as a first approach to semantic evaluation of alternative interpretations in NLP (Sadler 1989; Dagan et al. [(1991)]; Hindle 1990; Hindle, Rooth [(1991)]). Suppose, for instance, we wish to translate into English the Hebrew verb *lahtom* ('sign', 'seal', 'finish' or 'close'), in conjunction with the direct object *hoze* ('treaty'). Knowing the range of English verbs of which the word *treaty* can be the direct object can help to select *sign* as the most appropriate translation for the Hebrew verb (example from Dagan et al. *op. cit.*). Or consider the following job advertisement header:

(2) Freeze Dried Pharmaceuticals Manager

Corpus-based knowledge of words which can function as direct object to the verb *to dry* can provide the shallow kind of "knowledge of the world" which allows an NLP system to make the correct attachment of *Freeze Dried* to *Pharmaceuticals* rather than to *Manager*, without the need to introduce explicit slot restrictions, semantic features, etc.

Such disambiguation techniques require a large database of lexical relations. In many cases, however, direct lexical relations are insufficient for disambiguation. Consider, for example, the following pair of sentences.

(3) a. The man carrying the ladder broke the glass.
b. The man carrying the drinks broke the glass.

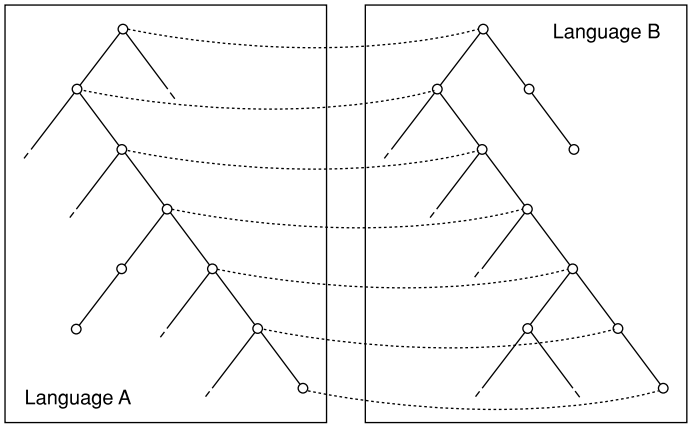


Figure 35.3
Coupling of two equivalent TKBs.

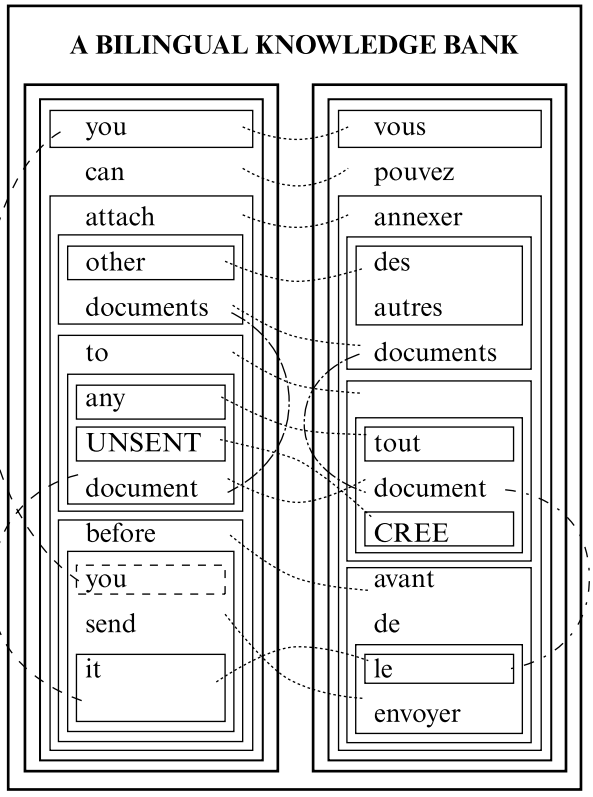


Figure 35.4
Fragment of a BKB.

In (3a), the most plausible interpretation of *glass* is 'window pane', whereas in (3b) the most likely meaning is 'wine or beer glass'. The interpretation of *glass* cannot be resolved only on the basis of its governor, the verb *broke*. The key to disambiguation in this example is to be found in the word *ladder* or *drinks*, which in a syntactic dependency representation has no direct relation to *glass* (see figure 35.5).

As the above example shows, an NLP system sometimes needs deeper or more extensive knowledge of conceptual, i.e., non- or extra-linguistic, relations than that provided by, for example, a database of lexical dependency relations. The knowledge base structure should make it possible to apply at least the simpler kinds of inference. So here again, just as for linguistic knowledge, there is a strong argument for using a textual knowledge bank containing not just dependency pairs, but complex structures corresponding to whole texts. If all the information contained in full text is preserved in the knowledge base, then in principle it should eventually be accessible to any NLP process, given appropriate software. This concept takes the principle of separating NLP programs from the linguistic data one step further. In TKB form, the data are separated not only from the program, but in principle from any application at all. It should be possible to build a TKB which can be used by many different kinds of application, perhaps based on quite diverse theoretical approaches, and to design NLP software which is entirely language-independent.

To sum up: the aim of building a Textual Knowledge Bank is to automate as far as possible the acquisition of the various types of knowledge required

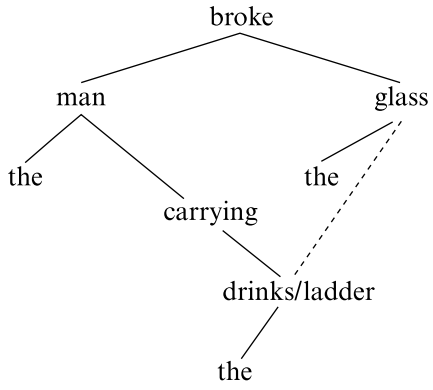


Figure 35.5
Indirect cues to disambiguation.

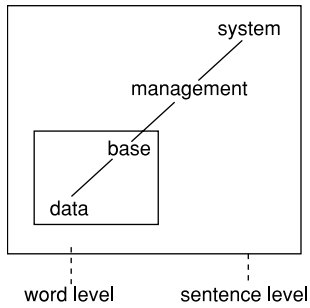


Figure 35.6
TKB relations on different levels.

for various forms of NLP—from monolingual knowledge of morphology, syntactic structures etc., through bilingual knowledge of lexical equivalences and translation syntax, to purely extra-linguistic knowledge of the world—by structuring the evidence which is explicitly and implicitly available in ordinary texts (and their translations).

TKB Design

The following description is drawn from prototype implementations currently under construction. It is intended primarily to give some idea of the possibilities. It is not intended to be in any way prescriptive. I shall not have much to say about the implementation as such. I will only point out that TKB implementations should allow fast access to a rather large number (typically millions) of word tokens in the corpus in order to permit efficient pattern matching between contextual patterns in the TKB and patterns in the input from a given application. They should also allow the original text to be reconstructed, either in its entirety or for any desired fragment. Beyond these general remarks, I shall now restrict myself to the conceptual design.

Language consists, in essence, of explicit, meaningful signs (typically words) and explicit or implicit relations between those signs (expressed by function morphemes, word order, intonation, punctuation or whatever). The starting point of our TKB design reflects this dichotomy: the TKB structure consists, basically, of objects and relations between them.

Objects include morphemes, words and multi-word units such as sentences and paragraphs. The whole TKB structure corresponding to a single coherent text can be viewed as a tree, with objects on the nodes, and relations on the branches of the tree. For languages such as English or Chinese, nodes are typically occupied by words; for agglutinative languages such as Inuit or Turkish, the objects on the nodes will more often be morphemes. Of course, any polymorphemic word is also an object in its own right, even if its constituent morphemes occupy several nodes in the tree. For example, the word *database* may be split across two tree nodes, containing the individual morphemes *data* and *base* respectively. Nevertheless, at the level of sentence structure, the whole word should also function as an object, with relations to other words in the text.

We have adopted the concept of level to distinguish between relations between morphemes, between words, and between sentences. For example, in the tree structure corresponding to the string *database management system* (figure 35.6), the node occupied by *base* has a dependency relation on one level with *data*, and on a different level with the node occupied by *management*. At this, the sentence level, the head node on the lower level represents its whole subtree (in this case, the whole word *database*).

The same principle applies to relations between sentences: these are represented by higher-level links between the head word node in one sentence and the head word node in the other sentence. Figure 35.2 provides an example, with a text-level relation drawn between the main verbs *use* and *can*.

Objects can be assigned a different form from that which appears literally in the text. This form is termed the normalized form. For example, the English word *gave* might be normalized to the form *give*. Figure 35.2 contains other examples, with normalization of plural to singular forms in *those* → that, *sharers* → sharer and *documents* → document. The aim of such normalization is to facilitate generalization over the TKB, e.g. for the purposes of semantic processing. At

the same time, the original surface form is also preserved and indexed, so that the same example (token) can be accessed either via the normalized form or via the surface form.

To each object, features can be attached if needed. For example, the word *give* with surface form *gave* might be assigned the feature "past tense".

Where normalization is regular and reversible, it can also be handled by rule-based processes at the input to and output from the TKB. For example, the English word *another* can be normalized to *a other* and reliably restored on the generation side. This is a type of normalization which is particularly useful where phonologically motivated changes are concerned.

The TKB requires a variety of relation types. Examples are: syntactic dependency, co-reference and bilingual equivalence. Other types can be introduced if required for a given application. Moreover, each relation can be labelled, e.g. with text grammatical or syntactic functions, such as subject, object, etc. Relations may be directed (e.g. governor-dependent) or bidirectional (e.g., bilingual equivalence).

It may be helpful to consider the various relation types under two headings—hierarchical and non-hierarchical.

Hierarchical relations can be divided into two types: syntactic and referential. In dependency syntax, each node in the syntactic tree has, in principle, only one governor or superordinate node (see for example the trees in figures 35.1–6). However, exceptions may be needed, especially in coordination. For instance, the word *British* in *British butterflies and moths* is a common dependent of both the coordinated nouns. For this reason, there is no true hierarchy in the syntactic structure, although for reasons of computational efficiency it may be decided to impose a true tree structure on the syntactic representation.

The second type of hierarchical relation appears in the identification of referential relations in the TKB corpus. For example, in

(4) Both parents must be held responsible; the father ...

the definite noun phrase *the father* refers back to the expression *both parents*. However, the reference type is not the usual one of identity, but one of inclusion. The concept 'parents' includes the concept 'father'. From such relations, a true hierarchy of (instances of) concepts can be built up across the TKB, and the usual consistency checks, transitivity mechanisms etc. can be applied to it.

Turning now to non-hierarchical relations, we can again distinguish two main types: referential and bilingual. All referential relations, such as the one illustrated by example (4), are basically monolingual (although they may be reflected in the parallel versions of the text in other languages). The commonest type of reference relation is that of identity. This can be exemplified by

(5) When you create a document you decide which folder will contain the document, and which drawer will contain the folder.

where *the document* is co-referent with *a document*, and *the folder* with *which folder*. We also distinguish a third type of reference relation, namely exclusion. Like identity (example 5), and unlike inclusion (example 4), exclusion too is non-hierarchical. This is the relation in

(6) Enter Y to delete the file; otherwise enter N.

between the adverb *otherwise* and the adjunct *to delete the file*, with which it is contrasted. (See also figure 35.2 for other examples of identity and exclusion.)

Bilingual relations are those which link one Textual Knowledge Bank (TKB) with another to form a Bilingual Knowledge Bank (or BKB). Since these links connect expressions which are declared to be equivalent in meaning, obviously they too are examples of non-hierarchical relations. Examples of bilingual relations can be seen in figure 35.4, where they are represented by dotted lines linking the English and French halves of a BKB. Each such link in figure 35.4 should be understood as equating two subtrees: the one governed by the node at the English end of the link, and the other governed by the node at the French end. For example, the link between the English word *to* and the French word *à* actually links the whole subtree (box) *to any UNSENT document* to the whole subtree *à tout document CRÉÉ*.

TKB Construction

The TKB concept has two different goals which, very fortunately, are in a way complementary. One aim is to provide a large linguistic and extra-linguistic knowledge base which can serve the purposes of analogical (or example-based) NLP. The other aim is to automate—at least to a considerable extent—the construction of such a knowledge base by taking machine-readable texts as the raw material of knowledge. Now once the first aim (of providing a basis for analogical NLP) is fulfilled, the second aim (building

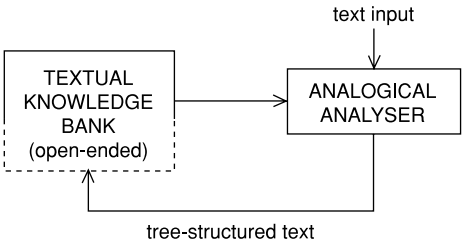


Figure 35.7
Self-improving TKB construction.

a knowledge base from raw text) can be achieved by designing and applying analogical tools. For example, a text can be converted to the TKB dependency tree structure by using an analogical dependency parser which uses a TKB as its knowledge source. The only trouble is that this is, of course, a chicken-and-egg situation.

Our answer has been to apply a kind of bootstrapping procedure, shown schematically in figure 35.7. Consider the parsing process for example. Starting with an empty TKB, the first sentence of the corpus text is parsed by hand and stored in the TKB. Thereafter, the system attempts to parse each new incoming sentence by retrieving subtrees whose projection matches part of the input string. (It may, of course, be supported by a few basic rules as well, if a hybrid approach is acceptable.) The provisional parse is corrected interactively and the corrected version is added to the TKB. (It is assumed that it is always possible to select one "correct" interpretation in cases of ambiguity.) Gradually, the models available in the TKB cover more and more of the possible grammatical constructions, and less and less intervention is demanded of the operator. Similar self-improving processes can be devised to perform morphological analysis, to add referential relations to a TKB, and to insert bilingual links between two TKBs in order to construct a Bilingual Knowledge Bank.

A major advantage of using TKB-based application software to construct a TKB in the first place, is that throughput from actual NLP applications can later be used to augment or supplement the existing knowledge bank. A machine translation system, for example, needs to build up in the course of translation a bilingual representation of the current text, complete with referential information. This representation of the source and target texts together is essentially equivalent to the BKB structure. A BKB-based MT system can therefore add its own output (after on-screen revision, if necessary) to its knowledge bank.

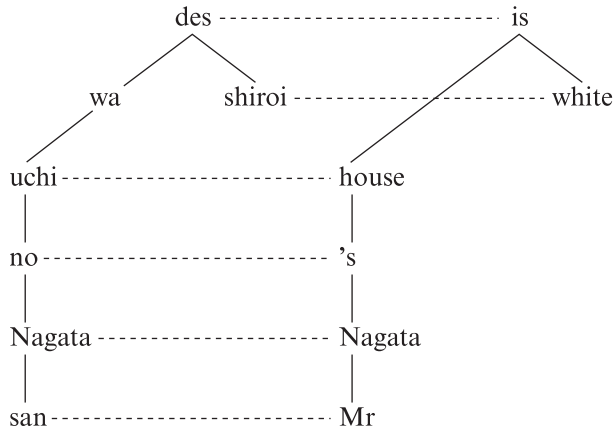


Figure 35.8
"Synsemizing" dependency trees.

This opens interesting perspectives for the automatic customization of NLP systems to user domains, vocabulary, style, etc.

The amount of interactive processing required in the early stages of TKB/BKB construction is such that an efficient user interface is essential. In our prototypes to date, the system's own attempts at analysis are presented graphically in the usual tree configuration (e.g., figure 35.8). The user can then edit the trees using a mouse.

Consider now the process of inserting bilingual links between two TKBs—what we have called "synsemizing" the trees, in other words aligning them semantically. Figure 35.8 shows a very simple example of synsemization between Japanese and English. Six bilingual relations have been identified. (Whether the identification bias been done interactively, or whether the system has identified all six relations on the basis of similar cases already stored in the BKB is not relevant here.) The interpretation of the bilingual relations is as follows. Each bilingual link between two nodes (words, morphemes or whatever) is by default interpreted to mean that the entire subtree governed by the node in one language is semantically equivalent to the entire subtree governed by the node in the other language. In other words, a bilingual link establishes a translation unit between two subtrees in different languages. In figure 35.8, for example, the link between *uchi* and *house* actually equates the whole subtree *Nagata san no uchi* with the whole subtree *Mr Nagata's house*.

Where, then, is the direct equivalence between *uchi* and *house* to be found? The answer is that this equivalence is derived. There is no need for the system, or

the operator building the BKB to identify this word-for-word equivalence, because it is implied by the principle of compositionality. Given (a) that the subtrees headed by *uchi* and *house* are equivalent, and (b) that the subtrees headed by the particles *no* and *'s* are equivalent, it follows (by the mechanism of tree subtraction) that *uchi* must be equivalent to *house*.

In this simple example, there is one element on the Japanese side—the particle *wa*—which has no bilingual relation with the English tree. The tree subtraction mechanism allows us to detach the subtrees headed by *uchi* = *house* and *shiroi* = *white*, producing the equivalence of *X wa Y desu* with *X is Y*.

Word order information is also preserved in the TKB data structure, though it is not visible in the figures shown here. Moreover, in our model, syntactic dependency relations (like all other relations in the database) can be labelled. In the figures shown here, syntactic labelling has been omitted for the sake of clarity, but it is important to understand that it is (or can be, depending on the application) part of the structures being synsemized.

A very important question in TKB construction is: To what extent should the lower-level structure of the language be analysed and represented in the trees? This question is very much tied up with the question of how far normalization of word forms should go. Ideally, all meaningful (translatable) elements in the language(s) concerned should occupy nodes in the tree structure, because only then can they be independently accessed and linked to other nodes, e.g., to their equivalents in another language. For example, given the Japanese–English equivalence *niwa-shi* = *gardener* should we be content to treat these expressions as units, occupying one node each? If so, it will not be possible to make use of the hidden equivalences *niwa* = *garden* and *-shi* = *-er*. Splitting the words across two nodes each will make the BKB more productive, but at the cost of expanding the data structure with additional nodes. There is a trade-off here, then, between productivity (lower-level analysis increases the power obtainable with a given corpus) and corpus size (without lower-level analysis, more examples will be needed to cover the same phenomena). The choice will often depend on the application. Thus TKB size depends on both corpus size and depth of analysis.

The interpretation of bilingual relations as referring to the entire subtrees governed by the nodes they link presupposes, of course, a certain degree of isomorphism between the dependency structures in the two languages being conjoined. As a default interpretation

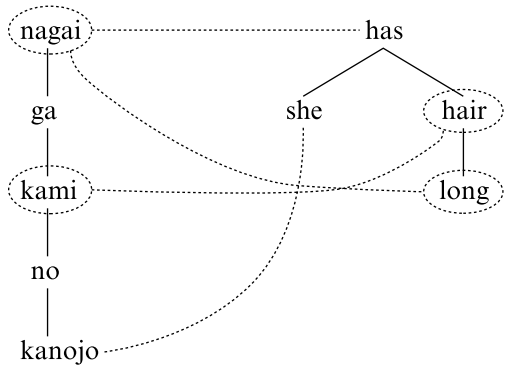


Figure 35.9
Heteromorphic constructions.

it is useful, but there are many cases where no such isomorphism exists. Figure 35.9 provides an example (from Watanabe 1990) where English and Japanese constructions are at odds. A bilingual link between *nagai* and *has* can be drawn, implying that the two expressions are equivalent as a whole. Also, a link can be drawn between *kanojo* and *she*, which have no dependents. But the sub-trees governed by *kami* = *hair* and *nagai* = *long* are not equivalent, although the words themselves are. In order to synsemize heteromorphic constructions such as this, an additional mechanism is required. Cases of heteromorphism such as this are quite common between Indo-European languages, although of course they may well be commoner between languages from different groups. Just as with the question of lower-level dependency analysis discussed above, there is a trade-off here between making a BKB fully productive (by applying an additional mechanism for heteromorphism), on the one hand, or making the data structure more complex (and thus increasing the storage space required), on the other.

It should be self-evident that a Bilingual Knowledge Bank, unlike most bilingual dictionaries, is fully reversible. There is no distinction in the data structure between "source" and "target" language. This is not to say that an equivalence such as *kami* = *hair* in figure 35.9 is universally applicable. There may be other English translations of *kami*, and other Japanese translations of *hair*. But the BKB structure clearly defines a context within which this equivalence obtains. Given that context, the equivalence is also reversible. Where a BKB is to be used as a machine dictionary, the reversibility principle means a very considerable saving as compared with conventional lexicographic methods.

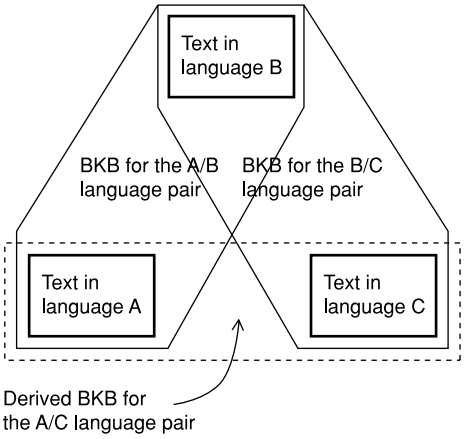


Figure 35.10
Deriving a BKB by transitivity.

There are further economies to be achieved where a multilingual NLP system is to be built. This is where the transitivity principle comes into play. Suppose we have already constructed two BKBs covering the same corpus in three different languages. In one, language A is coupled to language B; in the other, language B is coupled to language C. The two B halves are, of course, identical in that they contain the same text in the same language. Suppose further, that there arises a need for a direct translation system (or some other bilingual application) between languages A and C. To this end, a BKB has to be built which links these two languages (figure 35.10). Of course, this can be achieved using the same techniques already applied to produce the existing BKBs for the language pairs A/B and B/C. But much of the human understanding which went into building those BKBs was the same understanding now needed to construct the new BKB, because the same texts are involved. The question which therefore suggests itself is whether it is possible to extract from the existing BKBs the information required for the A/C combination. If so, then the new BKB can be compiled automatically.

As far as monolingual information structure is concerned, obviously no new work needs to be done. What remains to be done is to synsemize version A with version C by identifying the translation units (TUs) between them. Experiment with a pair of prototype BKBs (English–Esperanto and Esperanto–French) suggests that around 90% of the translation units can be automatically identified by the transitivity principle for this triangle of languages. The remainder can either be completed by the standard procedures or else compensated for by increasing the size of the corpus.

TKB Applications

The basic function of a Textual Knowledge Bank is to serve processes which attempt to match words and phrases from their input with the pre-analyzed examples in the TKB, in order to interpret their meaning and/or to check for consistency. This function suggests that TKB technology can form the basis for a wide variety of applications such as intelligent spelling and grammar checkers, extraction of technical terms, intelligent information retrieval and, in the longer term, machine translation. In this section I will consider a few of these applications by way of illustration. First, however, a few words about TKB applications in general.

A TKB can be used either as an off-line source for the extraction of statistical data, derivation of rules etc., or else as an on-line source for analogical processes. Using a TKB at runtime is motivated by the difficulty of defining in advance all the information which may possibly be needed. As an example, consider a semantic process that is able to compute a semantic probability on the basis of frequencies of words in context. Of course it would be theoretically possible to compute all semantic probabilities of this kind in advance, but the huge number of combinations in a reasonably large corpus advises against it. Instead, a powerful mechanism is needed at run time to derive the required information on the fly. Clearly, this choice between off-line and on-line usage involves questions of efficiency, storage etc. My remarks below on various applications will be primarily concerned with on-line usage, but this should not be taken to rule out off-line use or hybrid combinations of rule-based and analogical processes.

The very first applications which come to mind have already been mentioned above. I refer to the ''bootstrapping'' procedures for TKB construction described in the TKB Construction section. These were first developed for a pilot BKB implementation (Sadler and Vendelmans 1990) designed to explore the technology.

Now in order to serve as a general provider of linguistic and world knowledge, a TKB should contain large amounts of data. For time-critical TKB applications such as the BKB within an MT system, it is clear that efficient data storage techniques are needed. For this reason a small-scale implementation was designed which gave a good impression of a future large-scale system. For this pilot study, a small corpus of text from a software manual—some 20,000 words per language—was processed in English, French and Esperanto. From this corpus, three BKBs were built:

English/Esperanto, French/Esperanto and English/French. Viewed as a first *application* of TKB technology, the pilot implementation consisted of three main parts: the parser, the "synsemizer" and the retrieval system.

The parser which was used to parse the input text for the pilot implementation was a fast but simple, rule-based affair. The parsing process was only TKB-supported to the extent that information stored earlier was used to suggest word classes, features and normalized forms. Since then, several prototypes have been developed at BSO for a wholly TKB-based analogical parser which builds a dependency tree by comparing linear input patterns with the example trees in the TKB.

The pilot synsemizer was used very successfully to define translation units by establishing bilingual relations between corresponding monolingual subtrees. The system looked for probable translation units on the basis of the sentences already processed and displayed them for the operator's confirmation or correction. Subsequent proposals were influenced by the operator's response.

Lastly, the retrieval system is a tool which extracts information from a BKB that has been built using the parser and the synsemizer. On the basis of input phrases, which can include wild-cards and can be augmented with syntactic information, the BKB is queried. The resulting answers are presented to the user either graphically or textually. Possible queries include concordance queries, translation and back-translation queries, and—to some extent—bridge translation (e.g., simulated English-to-French translation via Esperanto by "chaining" two BKBS).

A much larger, multilingual TKB is now being built to revised linguistic and software specifications. The new implementation will also include a so-called "refalyzer" to attach anaphoric expressions to their antecedents and identify other kinds of reference relations.

Other rather basic applications for which prototype implementations already exist are the semantic proximity and semantic association functions. These are essential ingredients in any full-fledged analogical NPL system.

The computation of semantic proximity between different words or expressions in a given language is basic to the analogical, or example-based, approach. This is because any corpus is finite and some kind of extrapolation mechanism is needed to enable the system to evaluate input which cannot be literally matched with the knowledge base examples. Consider a simple example of reference ambiguity such as

(7) John took the cake from the table and washed it.

where the problem is to decide whether the pronoun refers back to the cake or to the table. If the knowledge base does not contain any examples of cakes or tables being washed, then semantic analogy is needed to make use of what examples are available (e.g. perhaps *eat cake* or *wash floor* or *clear table*). The question is: How similar is *eat* to *wash*, *floor* to *table*, *clear* to *wash*, etc.? We have implemented a semantic proximity function which answers this kind of question on the basis of a TKB by comparing the contextual patterns of the expressions concerned. This comparison is based, not on linear contexts in a corpus, but on the hierarchical context as defined by the dependency trees in the TKB. For example, using the English version of the software handbook TKB mentioned above, the semantic proximity function returns, for the input noun *file*, the following ranking of words with most similar contextual patterns:

0.101 director
0.053 document
0.044 format
0.033 mail
0.031 it
0.029 folder
0.027 message
0.023 reference

This constitutes a kind of dynamic thesaurus function in which the semantic relations are defined by the TKB corpus. One possible application is in speech recognition. Suppose an acoustic input has been tentatively deciphered as *edit the fi*[*lr*]*e* where there is ambiguity as to the last phoneme: [l] or [r]. Knowledge of contextual probabilities is needed to resolve the ambiguity. Taking the TKB model as the available knowledge base provides a number of known objects of the verb edit, such as *document*, *drawer*, *version*, *index*, etc., but no literal example in which either *file* or *fire* appears. The proximity function can be used to suggest the most likely cognates of the known examples, and the word *file* duly appears among the cognates of the commonest example, *document*.

The semantic association function is likewise based on contextual patterns, but returns expressions which

are syntagmatically, rather than paradigmatically, connected with the input expression. Thus, for example, the input word *file* will produce a list of terms such as *log*, *DOS*, *VMS*, etc., which all have a strong association with *file* but are not its cognates in the sense of being able to replace it in certain contexts.

This function has obvious applications in information retrieval. It can be used to prompt the database user with a list of associated terms to restrict or widen the search. Suppose, for example, a medical researcher enters the search term *serum*. On the basis of a very small medical TKB we have built, the user will then be offered such (syntactically) associated terms as *ferritin*, *IGF*, *protein*, *potassium*, etc. to use in combination with the input term. This is again a type of dynamic thesaurus function, but a different one from that offered by the semantic proximity function.

The Textual Knowledge Bank can also serve the purposes of error detection and correction, whether involving human error or defective transmission (e.g. in OCR, speech recognition etc.). Take the following example (from *New Scientist*, 15 July 1989):

(8) chemicals that inhabit the enzyme's activity

where the correct form should be *inhibit*. In this type of input error, there is no recognizable ungrammaticality, because both *inhabit* and *inhibit* are transitive verbs. The error can, however, be traced by a TKB-based semantic evaluator which recognizes the low plausibility of *inhabit* in this context. The system can then search for more plausible alternatives by exploring the known contextual relations in the TKB for each of the surrounding words in the input. Given a TKB with adequate biochemical coverage, it should not be difficult to trace the verb *inhibit* as a plausible governor of both *enzyme* and *activity*, allowing an orthographical proximity function to recognize the likelihood of a spelling mistake.

Speech recognition and synthesis could also be supported by a special type of TKB in which orthographic and phonetic representations are coupled to each other. Probabilistic techniques could then be used to evaluate alternative transcriptions in a context-sensitive fashion (Vendelmans 1989), where the TKB structure supports both linguistic and substantive analysis.

To conclude this brief account of possible applications for TKB technology, mention must also be made of bilingual applications such as machine translation and computer-aided translation, which, after all, were the original reason for devising the Bilingual Knowledge Bank. Some suggestions in this direction have already been made earlier in this paper (see also Sadler 1989 for more extended examples); moreover, BSO/Language Technology will shortly be embarking on further R&D concerning the mechanisms required for BKB-based translation.

References

Dagan, I., A. Itai, and U. Schwall. 1991. Two Languages Are More Informative Than One. *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, 130–137.

Hindle, D. 1990. Noun Classification from Predicate-Argument Structures. *28th Annual Meeting of the Association for Computational Linguistics*, Pittsburgh, 268–275.

Hindle, D., and M. Rooth. 1991. Structural Ambiguity and Lexical Relations. *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, 229–236.

Sadler, V. 1989. *Working with Analogical Semantics: Disambiguation Techniques in DLT*. Dordrecht: Foris.

Sadler, V., and R. Vendelmans. 1990. Pilot Implementation of a Bilingual Knowledge Bank. *13th International Conference on Computational Linguistics: Proceedings of COLING '90*, Helsinki, Vol. 3, 449–451.

Skousen, R. 1989. *Analogical Modeling of Language*. Dordrecht: Kluwer.

Vendelmans, R. 1989. A Structured Knowledge Bank for Syntactic and Semantic Speech Analysis. *Proceedings of the ESCA workshop on Speech Input/Output Assessment and Speech Databases*, Noordwijkerhout, The Netherlands, 5.11.1.

Watanabe, Hideo. 1990. A Model of a Transfer Process Using Combinations of Translation Rules. Submitted to PRICAI 1990.

Witkam, T. 1988. DLT—An Industrial R&D Project for Multilingual MT. *COLING Budapest: International Conference on Computational Linguistics*, 756–759.

36
Machine Translation without a Source Text

Harold L. Somers, Jun-ichi Tsujii and Danny Jones

Introduction

Machine Translation (MT) or natural language translation in general is a typical example of the "under-constrained" problems which we often encounter in the field of artificial intelligence[1]. That is to say, the same "messages" can and should be translated differently depending on the surrounding contexts (where and when they are used), and on the speakers' intention (what they really want to express), etc. It is all too often the case that this information, which is necessary for the selection of the appropriate overall target text structure, is not made explicit in source texts prepared for translation. The author of the source text naturally follows the "rules" of the source language in preparation of source texts and assumes that the factors which will affect the selection of target expressions are self-evident.

MT systems developed so far or being developed have been trying to compensate this genuine property of language translation by extending the units of translation from sentences to texts (e.g., Rothkegel 1986, Weber 1987) or by introducing "understanding" based on "domain specific knowledge" (as in the "sublanguage" approach—cf. Kosaka et al. 1988, Lehrberger and Bourbeau 1988). This course of research would be inevitable if we were to confine ourselves to translation of prepared texts which already exist before translation. In such cases, we have to recover from text itself or by using extra "knowledge", such implicit information which is necessary for formulating target expressions.

However, we can imagine a quite different course of research for developing a different type of MT system, i.e. an "expert" system which can play the role of an "intelligent secretary with knowledge of the foreign language". Such a system does not require the user (the writer) to prepare full source texts in advance. It starts from rough sketches of what the writer wants to say and gathers the information necessary for formulating target texts by asking the writer questions, because the writer is the person who really intends to communicate and has a clear idea about what s/he wants to say. We can get much richer information through such interactions than in the usual written text translation by professional translators. Through interaction, we can get information concerned with, for example, the user's intention which is not explicitly expressed in the "text" to translate but which is nonetheless necessary for producing quality target texts.

This sort of system is different from the widely promoted "Translator's Workbench" idea (e.g., Kay 1980, Melby 1982), the main aims of which are to help translators to translate texts. In this scenario, both the system and the user have knowledge about both source and target language, and it is sometimes difficult to see where the most appropriate division of labour should occur: indeed, there is sometimes a conflict between what the system offers the translator–user, and what the user already knows, or between the extent to which the system or the user should take the initiative, which might differ from occasion to occasion. On the other hand, in the proposed expert system scenario, the partition of knowledge is clear: the system knows mainly about translation, the writer knows only about the desired communicative content of the message. There is no conflict between what the system assumes to be the extent of the writer's (the user's) knowledge, nor in the writer's expectations. In this respect we are following the line taken by Johnson and Whitelock (1987), and the work here at UMIST on the ENtran project (Whitelock et al. 1986, Wood and Chandler 1988) developing an MT system for a monolingual user.

MT systems so far have been developed based on the implicit assumption that source texts contain all (or almost all) the information necessary for translation. We take as a starting point that this assumption is not necessarily true, especially when we consider pairs of unrelated languages where cultural as well as linguistic differences contribute to this problem.

Notice that the concept of "source text" in the above is quite different from that in the normal

context of MT. That is, we do not have a source text to translate as such, but instead, the user has his/her communicative goals and the translation system can help to formulate the most appropriate target linguistic forms by gathering information necessary to accomplish these goals through "clarification dialogues".

It could be argued that this generation of a target text on the basis of something other than a source text is not "real translation". Such an argument might derive from an overly traditional view of translation where a translator gets some text (say, in the post) and sits at a desk with a bilingual dictionary and translates "blind", i.e., with no actual knowledge of the writer's intentions, goals, etc. There is a sense in which second generation MT systems simply reflect this scenario of a translator. Of course, the best translations are done by a translator who can ask the original author "What did you mean when you said ...?"; by the same token we believe we can build a better translation system if we can elicit such information from the originator of the "text" at the time of "writing".

General Background to the Research

This research is undertaken in the context of the more general activities of the Japanese ATR research programme into automatic interpretation between English and Japanese of telephone conversations. As such it is oriented towards translation of dialogues. One approach to dialogue translation has been the "phrasebook" approach of Steer and Stentiford (1989). In this speech translation prototype system, set phrases are stored, as in a holidaymaker's phrasebook; they are retrieved by the fairly crude, though effective, technique of recognizing keywords in a particular order in the input speech signal. The main disadvantage of this system is its inflexibility: if the phrase you want is not in the phrasebook, you cannot say anything.

In the research programme to be reported here, we are not concerned with speech processing per se, and we assume the context of an on-line keyboard conversation function such as `talk` in UNIX (cf. Miike et al. 1988). It has been found that keyboard conversations have the same fundamental features as telephone conversations, notwithstanding the obvious differences between written and spoken language (Arita et al. 1987, Iida 1987). Furthermore, we restrict ourselves to goal-oriented dialogues, i.e., dialogues where one participant is seeking information from the other: our experimental domain is dialogues for a conference registration and hotel reservation system.

When such conversations are subjected to the additional distortion of being transmitted via a traditional MT system, several further problems accrue, as the `talk` experiment mentioned above showed, notably when mistranslation occurs. The problem of human–machine interaction in the specific area of clarification dialogues for MT must be studied. The need to incorporate different types of clarification dialogue has general implications for the question of system architectures for interactive MT systems. This aspect is discussed in detail below.

In the above scenario, the system tries to gather information necessary for formulating target texts through interactions. This means the system formulates target texts by adding information to "source texts" (in the conventional sense). We can extend this idea further. In the extreme case, we can imagine a system which has stereotypical target texts in certain restricted domains (e.g., business correspondences in specific areas), retrieves appropriate texts through dialogues with users and reformulates them to fulfill the specific requirements expressed by users. In this scenario, the MT system becomes a kind of multilingual text generation system and adds a lot of information not contained in the "source text" at all. This idea has been investigated here at UMIST in the context of a research programme for British Telecom (Jones and Tsujii 1990), and has significantly influenced the research reported here (a similar idea for "automated text composition" in Japanese has been suggested by Saito and Tomita 1986).

Dialogue MT

It is important to emphasize that there is a basic difference between Dialogue Machine Translation (DMT) systems on the one hand and conventional MT systems on the other, namely the difference of user types. In DMT, users are dialogue participants who actually have their respective communicative goals and who really know what they want to say. On the other hand, the users of conventional MT are typically translators who, though they have enough knowledge about both languages, lack "complete understanding" of texts to be translated.

This difference in user-types leads to different characterizations of interactions between MT systems and their users. We have to take into account what this difference implies in designing actual DMT systems. The main implications can be summarized as follows.

In DMT, the system can ask in theory any questions to elicit the information necessary for translation which is not explicitly expressed in the "source text". This is impossible in conventional MT, because the users do not have "complete understanding" of the context in which the texts are prepared, and the users (who are translators) simply could not answer such questions. (It is often the case that even human translators would like to consult the authors of the original texts in order to produce a good translation.) In order to exploit this advantage in DMT however, we have to overcome several related difficulties.

First, in DMT there are several different types of dialogues, any of which may start up or be resolved at any given time: these dialogues include

a. user–user object-level dialogues

b. user–user meta-level dialogues (e.g. in which one participant in the dialogue asks the other participant questions to clarify the meaning or intentions of his/her statements)

c. user–system dialogues typically initiated by the system, concerning the progress of the object-level dialogue, disambiguating ambiguous object-level dialogue, i.e., what the user wants to *say next*

d. user–system meta-level dialogues typically initiated by the user, concerning clarification of the object-level dialogue, i.e., what was *just said*

One of the foreseeable difficulties in DMT is how to distinguish these different modes of dialogue, that is, how systems can distinguish, first of all, utterances of types (a) and (b) to be translated and transmitted, from utterances of type (d) which should not be translated. In particular, dialogues of types (b) and (d) may be difficult in some cases, because the user posing questions of clarification cannot generally recognize whether the difficulties of understanding come from "errors" in translation or from the other participants' utterances themselves. For examples of this effect, see Miike et al. (1988).

Dialogues of type (c) are found in some form in most conventional interactive MT systems; note that with monolingual users such dialogues are quite different from those found in the "Translator's Workbench" type of system, since it is particularly difficult to phrase interactions concerning problems of transfer when the user is not expected to know anything about the target language, and when current frameworks do not allow us to specify the relationships among possible translations defined by different structural correspondence rules. On the other hand, regarding problems with analysis, a particularly useful result of the research on ENtran was to see to what extent potential ambiguities could be recognized on the basis of structures computed by more or less traditional parsing techniques (i.e. charts). For dialogues of type (c) we are guided by the work of Jones and Tsujii, mentioned above.

The British Telecom work concerns a system for generating business letters in French, German, and Spanish on the basis of an essentially menu-driven interface (in English). The system has a set of pretranslated fragment pairs some of which have slots for variable elements to be inserted (e.g., the name of a company, or a product) which may or may not be translated in a conventional manner. The system–user dialogue aims at selecting the appropriate target-language expression (TLE) fragment corresponding to some source-language expression (SLE) and compiling the TLEs in the appropriate sequence so as to generate the required output. Notice that, since the fragments have been pretranslated (presumably by a competent translator), the result is of a guaranteed high quality.

This idea is developed in the following ways. First, we assume that the interface menu is replaced by a much more complex "model dialogue" (see below). In the sense that the pretranslated fragment pairs are associated with particular points in the model dialogue, they can be said to be not just pairs of SLEs and TLEs but in fact triples, since they are identified by a description of the dialogue context (DC) which conditions the equivalence of the two expressions, by specifying the point in the model dialogue at which they are identified, thus: $\langle \mathrm{SLE},\mathrm{TLE},\mathrm{DC} \rangle$. It is possible for a given SLE, there may be several TLEs depending on the particular DC, thus:

$\langle \mathrm{SLE}_m,\mathrm{TLE}_i,\mathrm{DC}_x \rangle$
$\langle \mathrm{SLE}_m,\mathrm{TLE}_j,\mathrm{DC}_y \rangle$
$\langle \mathrm{SLE}_m,\mathrm{TLE}_k,\mathrm{DC}_z \rangle$

For example, the English response *OK* in a dialogue may correspond to Japanese *wakarimashita* when something is being explained, *ii desu yo* when asserting agreement, or *ijō desu* when it indicates completion of the discussion and a change of topic.

The task of the DMT system can now be divided between first locating the appropriate set of triples involving a given SLE, and then locating the appropriate TLE for that SLE according to the DC.

If we assume that the SLEs are not just "canned texts", but actually types of text templates of varying linguistic complexity (i.e., from set phrases through to

syntactic patterns—see below), it can be seen that the first part of the above task can be achieved by traditional techniques of parsing or by some other matching procedure. The set of different DCs for a given SLE can be used to trigger a clarification dialogue so as to determine the appropriate TLE.

In this scenario the user has taken the initiative in the dialogue, by "typing in" what s/he wants to say, and having the system find the appropriate triple.

Two other scenarios are also possible. In one, the system retains the initiative, and rather like in the menu-driven system, selects (or seeks via a meta-dialogue) the next appropriate DC, and then offers a range of appropriate SLEs for selection. In this sense the ⟨SLE,TLE⟩ pair for a given value of DC can be regarded as a "conditioned equivalence pair".

Finally, in a mixed-initiative scenario, the user and the system collaborate in the following way: first, a communicative goal is established, and with it a sequence of DCs corresponding to the "dialogue plan". The user then makes a proposal for the next utterance in the dialogue, and the system searches its database for the nearest apparently appropriate ⟨SLE,TLE,DC⟩ given the user's input (corresponding to the SLE) and the DC as given by the dialogue plan. If an exact match is found, the TLE is generated and the object-level dialogue continues. However, if an exact match is not found, the system gets the user to modify the SLE until it more closely matches the SLE selected by the system.

Model Dialogue

The important issue in the above is that the equivalence relation of the two expressions is not guaranteed by the expressions themselves but by the DCs which are given rather independently of the informational content of the two expressions in the triples. In a context such as business correspondence, it might be the case that much less information is necessary to identify the relevant triple than that conveyed by the actual linguistic expressions and that, because each individual language usually has its own conventions which letters must follow, the actual informational contents of the two expressions might be different. The same is true of certain types of dialogues. For example, there are conventional phrases used in Japanese phone calls (Nagasaki 1971) which, if translated literally, would probably mystify the non-Japanese dialogue partner:

- *Sorry to disturb you when you are busy|eating|about to go to bed|still asleep* (depending on time of day)
- *Sorry to have had to disturb you*
- *Sorry for having talked too much*
- *Excuse me for bothering you*
- *Thank you for going out of your way to answer the phone*
- *I assume it is inconvenient for you now, but ...*
- *I am sorry for phoning you without warning*
- *I wasn't expecting to phone you, but ...*

One important research question is what exactly the DC should look like. Our current assumption is that DC will actually refer to a point in a "model dialogue", probably a flexible network of script-like structures indicating possible dialogues that the system can translate, perhaps along the lines of work by Wachtel (1986) and Reilly (1989). We have not yet finalized our ideas in this area, but we are considering in particular how to model suitably flexible dialogue structures within the domain in question, the problem of interactions between the model dialogues and the meta-dialogues, as well as the mechanisms which enable the system to navigate its way through the model dialogue network in response to the user's input.

"Canned Text" and Extensions

It was stated above that the nature of the SLE and TLE pairs should be varied. In particular, because of the need for flexibility as compared to the British Telecom work described in Jones and Tsujii (1990), we assume that the system will permit some degree of conventional compositional translation. So SLEs and TLEs are not always texts, or "paratexts" (i.e., texts with slots for proper names or simply translated noun phrases, etc.) but, in some cases, structural descriptions of a more conventional kind. This clearly implies that within the system there is a need for analysis (and generation) of the kind found in conventional MT systems. In particular, where appropriate texts or paratexts are not found for a given input, and the dialogue management part of the system is satisfied that "free input" is an available option at this point in the model dialogue, then the system becomes more like a conventional MT system, though with the special characteristics of an MT system which interacts with a monolingual user.

For the most part, however, it is assumed that there is a stereotyped set of functions involved in perform-

ing a global communicative function in a restricted domain. We can assign surface representations to these functions which restrict the form of expression to a certain extent in order to capture functional regularities in communication and to guarantee high quality translations. When the system encounters unexpected input, it has a choice of trying to steer the user towards input which is more within its expectations, or to abandon temporarily its assurance of high-quality translation and operate in a more traditional manner.

It may be asked why we need the model dialogues, the canned text and paratexts, and conditioned equivalence pairs: would it not be better simply to have a long pre-composition phase where the writer interacts with an expert system which asks lots of questions about intentions and goals and then uses this knowledge (if required) in a conventional parse-and-disambiguate system? Of course this would be another way of addressing the problem of underspecified texts, but it is not clear what type of questions could be asked unless a specific domain of composition was pin-pointed. This brings us back to domain knowledge, which in this case is expressed as knowledge about what the user can ask next, which we capture in the model dialogues.

Conclusion

It is nowadays accepted that we cannot expect to have fully automatic high-quality MT. We have to develop systems which allow flexible and effective human inteventions. Our idea is to explore diversified approaches to interactive MT and in particular we seek to develop an interactive system for monolingual users. Furthermore, it seems that several interesting new approaches become apparent once we escape from the basic assumption of the existence of a concrete source text, and explore the idea of "MT without source texts".

Notes

1. The authors would like to acknowledge the contribution to this work of the other members of the project team: Bill Black, Jeremy Carroll, Anna Gianetti, Makoto Hirai, Natsuko Holden, John Phillips, and Kenji Yoshimura.

2. Our concept of DMT should be distinguished from "Dialogue-*based* MT" as proposed by Boitet (1989), in which dialogue is used to clarify the author's intentions in the context of a personal MT system. This is also the case in *our* DMT, with the crucial difference that the object of translation in our case is also part of a dialogue, i.e., the user's dialogue with a third party. Clearly, however, there are significant areas of overlap between our project and Boitet's.

References

Arita, H., K. Kogure, I. Nogaito, H. Maeda, and H. Iida. 1987. Media ni izon suru kaiwa no yōshiki: denwakaiwa to kiibōdo no kaiwai no hikaku (Media-dependent conversation manners: comparison of telephone and keyboard conversations). *Jōhō Shori Gakkai* 87.34 Jōhō Shori Gakkai Kenkyū Hōkoku, Shizen Gengo Shori 61-NLP-5, 1987.5.22.

Boitet, C. 1989. Speech Synthesis and Dialogue Based Machine Translation. *ATR Symposium on Basic Research for Telephone Interpretation*, Kyoto, December 1989. Preprints, 6-5-1-9.

Iida, H. 1987. Distinctive Features of Conversations and Interkeyboard Interpretation. Paper presented at Workshop on Natural Language Dialogue Interpretation, Advanced Telecommunications Research Institute (ATR), Osaka, November 1987.

Johnson, R. L., and P. Whitelock. 1987. Machine Translation as an Expert Task. In S. Nirenburg (ed.), *Machine Translation: Theoretical and Methodological Issues*. Cambridge: Cambridge University Press, 136–144. Reprinted in this collection.

Jones, D., and J. Tsujii. 1990. High Quality Machine Translation for Monolinguals. *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, Austin, Texas, 43–46.

Kay, M. 1980. The Proper Place of Men and Machines in Language Translation. Research Report CSL-80-11. Xerox Palo Alto Research Center, Palo Alto, California, October 1980. Reprinted in this collection.

Kosaka, M., V. Teller, and R. Grishman. 1988. A Sublanguage Approach to Japanese-English Machine Translation. In D. Maxwell, K. Schubert, and T. Witkam (eds.), *New Directions in Machine Translation*. Dordrecht: Foris, 109–120.

Lehrberger, J., and L. Bourbeau. 1988. *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation*. Amsterdam: John Benjamins.

Melby, A. K. 1982. Multi-level Translation Aids in a Distributed System. In J. Horecký (ed.), *COLING 82: Proceedings of the Ninth International Conference on Computational Linguistics*. Amsterdam: North-Holland, 215–220. Reprinted in this collection.

Miike, S. K. Hasebe, H. Somers, and S. Amano. 1988. Experiences with an On-line Translating Dialogue System. *26th Annual Meeting of the Association for Computational Linguistics*, Buffalo, NY, 155–162.

Nagasaki, K. 1971. *(Hito ni warawarenai) Kotoba dzukai to hanashi kata*. Tōkyō: Bunwa Shobo.

Reilly, R. 1989. Communication Failure in Dialogue: Implications for Natural Language Understanding. In J. Peckham (ed.), *Recent Developments and Applications of Natural Language Processing*. London: Kogan Page, 244–261.

Rothkegel, A. 1986. Textverstehen und Transfer in der maschinellen Übersetzung. In I. Bátori and H. J. Weber (Hgg), *Neue*

Ansätze in Maschineller Übersetzung: Wissensrepräsentation und Textbezug. Tübingen: Max Niemeyer Verlag, 197–227.

Saito, H., and M. Tomita. 1986. On Automatic Composition of Stereotypic Documents in Foreign Languages. Presented at 1st International Conference on Applications of Artificial Intelligence to Engineering Problems, Southampton, April 1986. Research Report CMU-CS-86-107, Department of Computer Science, Carnegie Mellon University.

Steer, M. G., and F. W. M. Stentiford. 1989. Speech Language Translation. In J. Peckham (ed.), *Recent Developments and Applications of Natural Language Processing*. London: Kogan Page, 129–140.

Wachtel, T. 1986. Pragmatic Sensitivity in NL Interfaces and the Structure of Conversation. *11th International Conference on Computational Linguistics, Proceedings of COLING-86*, Bonn, 35–41.

Weber, H. J. 1987. Converging Approaches in Machine Translation: Domain Knowledge and Discours [sic] Knowledge. Linguistic Agency University of Duisburg Series B, No. 164.

Whitelock, P. J., M. M. Wood, B. J. Chandler, N. Holden, and H. J. Horsfall. 1986. Strategies for Interactive Machine Translation: The Experience and Implications of the UMIST Japanese Project. *11th International Conference on Computational Linguistics, Proceedings of COLING-86*, Bonn, 329–334.

Wood, M. M., and B. J. Chandler. 1988. Machine Translation for Monolinguals. In D. Vargha (ed.), *COLING Budapest: Proceedings of the 12th International Conference on Computational Linguistics*. Budapest: John von Neumann Society for Computing Sciences, 760–763.

Source Notes

We have made every effort to ascertain and acknowledge ownership of copyright for all the articles reprinted in this collection. Where not stated, copyright is with the original author, or unknown. The editors would be grateful to hear of any corrections to this list.

"Translation" by Warren Weaver and "The Mechanical Determination of Meaning" by Erwin Reifler both appeared in William N. Locke and A. D. Booth (eds.), *Machine Translation of Languages: Fourteen Essays*, jointly published by The Technology Press of the Massachusetts Institute of Technology, John Wiley (New York), and Chapman & Hall (London). Copyright © 1955 The Massachusetts Institute of Technology and reprinted with kind permission of The MIT Press.

"Mechanical Translation" by A. D. Booth appeared in *Computers and Automation*, vol. 2, no. 4, in 1955.

The following articles appeared in the journal *Mechanical Translation*, copyright © The MIT Press: "Stochastic Methods of Mechanical Translation" by Gilbert W. King (volume 3, 1956), "A Framework for Syntactic Translation" by Victor H. Yngve (volume 4, 1957), and "A New Approach to the Mechanical Syntactic Analysis of Russian" by Ida Rhodes (volume 6, 1961).

"The Present Status of Automatic Translation of Languages" by Yehoshua Bar-Hillel appeared in Franz L. Alt (ed.), *Advances in Computers*, Vol. 1, Academic Press (New York), copyright © 1960 Academic Press.

"A Preliminary Approach to Japanese–English Automatic Translation" by Susumu Kuno and "On the Mechanization of Syntactic Analysis" by Sydney M. Lamb appeared in *1961 International Conference on Machine Translation of Languages and Applied Language Analysis: Proceedings of the Conference held at the National Physical Laboratory, Teddington, Middlesex, on 5th, 6th, 7th and 8th September* (National Physical Laboratory Symposium No. 13), published by Her Majesty's Stationery Office (London), 1962.

"Research Procedures in Machine Translation" by David G. Hays appeared in Paul L. Garvin (ed.), *Natural Language and the Computer* (University of California Engineering and Sciences Extension Series), published by McGraw-Hill (New York), 1963.

"ALPAC: The (In)Famous Report" by John Hutchins appeared in *MT News International*, no. 14 (June, 1996), and is reprinted by kind permission of the author.

The following papers appeared in A. D. Booth (ed.), *Machine Translation*, published by North-Holland Publishing Company (Amsterdam), 1967: "Correlational Analysis and Mechanical Translation" by Silvio Ceccato, "Automatic Translation: Some Theoretical Aspects and the Design of a Translation System" by O. S. Kulagina and I. A. Mel'čuk, and "Mechanical Pidgin Translation: An Estimate of the Research Value of 'Word-for-Word' Translation into a Pidgin Language, Rather Than into the Full Normal Form of an Output Language" by Margaret Masterman.

"English–Japanese Machine Translation" by S. Takahashi, H. Wada, R. Tadenuma, and S. Watanabe appeared in *Information Processing*, *Proceedings of the International Conference on Information Processing*, UNESCO, Paris, 15–20 June 1959, jointly published by R. Oldenbourg (München) and Butterworths (London).

"Computer Aided Translation: A Business Viewpoint" by John S. G. Elliston appeared in Barbara M. Snell (ed.), *Translating and the Computer: Proceedings of a Seminar*, London, 14th November 1978, published by North-Holland (Amsterdam), 1979.

"Automatic Translation and the Concept of Sublanguage" by J. Lehrberger appeared in R. I. Kittredge and J. Lehrberger (eds.), *Sublanguage: Studies of Language in Restricted Semantic Domains*, published

by Mouton de Gruyter (Berlin) in 1982, and is reprinted by kind permission of the author.

"The Proper Place of Men and Machines in Language Translation" by Martin Kay first appeared as Research Report CSL-80-11, Xerox Palo Alto Research Center, Palo Alto, California, copyright © 1980 Xerox Corporation. Since permission to include it in this collection was obtained, it has also been reprinted in *Machine Translation*, vol. 12 (1997).

"Machine Translation as an Expert Task" by Roderick L. Johnson and Peter Whitelock appeared in S. Nirenberg (ed.), *Machine Translation: Theoretical and Methodological Issues*, copyright © 1985 Cambridge University Press, and is reprinted with the permission of Cambridge University Press.

"Montague Grammar and Machine Translation" by Jan Landsbergen appeared in P. Whitelock, M. M. Wood, H. L. Somers, R. Johnson, and P. Bennett (eds.), *Linguistic Theory and Computer Applications*, published by Academic Press (London) in 1987.

The following papers from COLING conferences are copyright © ICCL and are reprinted with their kind permission: "Automatic Translation—A Survey of Different Approaches" by B. Vauquois from COLING-76 (Ottawa), "Dialogue Translation vs. Text Translation—Interpretation Based Approach" by Jun-ichi Tsujii and Makoto Nagao from COLING-88 (Budapest), and "Machine Translation without a Source Text" by Harold L. Somers, Jun-ichi Tsujii, and Danny Jones from COLING-90 (Helsinki).

"Translation by Structural Correspondences" by Ronald M. Kaplan, Klaus Netter, Jürgen Wedekind, and Annie Zaenen appeared in *Proceedings of the Fourth Conference of the European Chapter of the Association for Computational Linguistics*, Manchester, 1989. Reprinted with the kind permission of the Association for Computational Linguistics.

"Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation" by Christian Boitet appeared in Dan Maxwell, Klaus Schubert, and Toon Witkam (eds.), *New Directions in Machine Translation* (Distributed Language Translation 4), published by Foris (Dordrecht) in 1988, and is reprinted by permission of Mouton de Gruyter, a division of Walter de Gruyter & Co., Berlin.

"Treatment of Meaning in MT Systems" by Sergei Nirenburg and Kenneth Goodman appeared in *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language* (Linguistics Research Center, University of Texas at Austin), copyright © 1990 University of Texas at Austin.

"Where Am I Coming From: The Reversibility of Analysis and Generation in Natural Language Processing" by Yorick Wilks appeared in Martin Pütz (ed.), *Thirty Years of Linguistic Evolution*, copyright © 1990 John Benjamins Publishing Company, Philadelphia/Amsterdam.

"The Place of Heuristics in the Fulcrum Approach to Machine Translation" by Paul L. Garvin appeared in *Lingua*, vol. 21 (1968).

"Three Levels of Linguistic Analysis in Machine Translation" by Michael Zarechnak appeared in *Journal of the Association for Computing Machinery*, vol. 6 (1959), copyright © 1959 ACM, Inc., reprinted with permission.

"Multi-level Translation Aids" by Alan K. Melby appeared in J. Horecky (ed.), *COLING-82: Proceedings of the Ninth International Conference on Computational Linguistics* (North-Holland Linguistic Series 47), published by North-Holland (Amsterdam), 1982. Copyright © Academia.

The extract from "EUROTRA: Computational Techniques" by Rod Johnson, Maghi King, and Louis des Tombe appeared in *Computational Linguistics*, vol. 11, copyright © 1985 the Association for Computational Linguistics, and is reprinted with permission.

"A Framework of a Mechanical Translation between Japanese and English by Analogy Principle" by Makoto Nagao appeared in Alick Elithorn and Ranan Banerji (eds.), *Artificial and Human Intelligence* (edited review papers presented at the International NATO Symposium on Artificial and Human Intelligence sponsored by the Special Programme Panel held in Lyon, France, October, 1981) published by North-Holland (Amsterdam) in 1984, and is reprinted with the kind permission of the author.

"A Statistical Approach to Machine Translation" by Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin appeared in *Computational Linguistics*, vol. 16, copyright © 1990 the Association for Computational Linguistics, and is reprinted with permission.

"Automatic Speech Translation at ATR" by Tsuyoshi Morimoto and Akira Kurematsu appeared in *The Fourth Machine Translation Summit Proceedings, International Cooperation for Global Communication* (Kobe, Japan), copyright © 1993 Steering Committee of the Fourth Machine Translation Summit, and is reprinted with permission.

"The Stanford Machine Translation Project" by Yorick Wilks appeared in R. Rustin (ed.), *Natural Language Processing* (Courant Computer Science Symposium 8, December 1971) published by Algorithmics Press (New York) in 1973.

"The Textual Knowledge Bank: Design, Construction, Applications" by Victor Sadler was presented at the International Workshop on Fundamental Research for the Future Generation of Natural Language Processing (FGNLP), 23–24 July 1991, Kyoto, Japan, organized by ATR Interpreting Telephony Research Laboratories.

This page intentionally left blank

Index