In the Name of the Most High

A Rapid Review

Focus on MACHINE TRANSLATION

(2)

Essential Notes and Tests

For M.A. & PhD Candidates

Including:

Machine Translation: An Introductory Guide (D. Arnold, 1996)
Speech and Language Processing (D. Jurafsky, & J. H. Martin, 2007)
Machine Translation: Its Scope and Limits (Y. Wilks, 2009)
Computers and Translation: A translator's guide (H. Somers, 2003)
Learning Machine Translation (C. Goutte, 2009)
Introducing Electronic Text Analysis (S. Adolphs, 2006)

Mahmoud Ordudari
PhD Candidate of Translation Studies
University of Allameh Tabatabai

Hussein Mollanazar (PhD)
Assistant Professor
University of Allameh Tabatabai

2012

Contents

مقدمه

کتاب حاضر دومین جلد از مجموعه کتاب‌های «رایانه و ترجمه» بوده و به عنوان کتاب کمک درسی برای واحدی با همین نام در دوره کارشناسی ارشد و دکتری مطالعات ترجمه تهیه و تدوین شده است. بعلاوه، کتاب می‌تواند برای دانشجویان کامپیوتر و هوش مصنوعی نیز بسیار مفید باشد.

کتاب به شیوه «مرور سریع» (*Rapid Review*) دربرگیرنده‌ی نکته‌های مهم شش کتاب مفید در زمینه ترجمه ماشینی است:

- Machine Translation: An Introductory Guide (Arnold, Balkan, Meijer, Humphreys & Sadler, 1996)
- Speech and Language Processing (D. Jurafsky, & J. H. Martin, 2007)
- Machine Translation: Its Scope and Limits (Y. Wilks, 2009)
- Computers and Translation: A translator's guide (H. Somers, 2003)
- Learning Machine Translation (Goutte, Cancedda, Dymetman, & Foster, 2009)
- Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies (S. Adolphs, 2006)

برای استفاده بهینه از این کتاب، پیشنهاد می‌شود در ابتدا کتاب اصلی به دقت خوانده شود و سپس برای تقویت یادگیری مطالب به نکته‌ها و در نهایت برای ارزیابی میزان یادگیری، به تست‌ها مراجعه شود.

Book ❶

Machine Translation:

An Introductory Guide

Douglas Arnold / Lorna Balkan / Siety Meijer
R. Lee Humphreys / Louisa Sadler

1.1 Notes

Chapter 1
Introduction and Overview

✍ **1. Scientifically**, MT is interesting, because it is an obvious application and testing ground for many ideas in Computer Science, Artificial Intelligence, and Linguistics, and some of the most important developments in these fields have begun in MT.

✍ **2. Philosophically**, MT is interesting, because it represents an attempt to automate an activity that can require the full range of human knowledge—that is, for any piece of human knowledge, it is possible to think of a context where the knowledge is required.

✍ **3.** The criticism that MT systems cannot, and will never, produce translations of great literature of any great merit is probably correct, but quite beside the point. It certainly does not show that MT is impossible.

✪ First, translating literature requires special literary skill—it is not the kind of thing that the average professional translator normally attempts. So accepting the criticism does not show that automatic translation of non-literary texts is impossible.

※ Second, literary translation is a small proportion of the translation that has to be done, so accepting the criticism does not mean that MT is useless.

◈ Finally, one may wonder who would ever want to translate Shakespeare by machine—it is a job that human translators find challenging and rewarding, and it is not a job that MT systems have been designed for.

✍ **4.** The **quality** of translation that is currently possible with MT is one reason why it is wrong to think of **MT systems** as *dehumanizing monsters* which will eliminate human translators, or enslave them. It will

not eliminate them, simply because the volume of translation to be performed is so ***huge***, and ***constantly growing***, and because of the limitations of current and forseeable MT systems. While not an immediate prospect, it could, of course, turn out that MT enslaves human translators, by controlling the translation process, and forcing them to work on the problems it throws up, at its speed.

✍ **5.** Some **Facts about MT:**

- ✵ MT is useful. The METEO system has been in daily use since 1977. As of 1990, it was regularly translating around 45000 words daily. In the 1980s, The diesel engine manufacturers Perkins Engines was saving around £4000 and up to 15 weeks on each manual translated.
- ☆ While MT systems sometimes produce howlers, there are many situations where the ability of MT systems to produce reliable, if less than perfect, translations at high speed is valuable.
- ❄ In some circumstances, MT systems can produce good quality output: less than 4% of METEO output requires any correction by human translators at all (and most of these are due to transmission errors in the original texts). Even where the quality is lower, it is often easier and cheaper to revise 'draft quality' MT output than to translate entirely by hand.
- ⌖ MT does not threaten translators' jobs. The need for translation is vast and unlikely to diminish, and the limitations of current MT systems are too great. However, MT systems can take over some of the boring, repetitive translation jobs and allow human translation to concentrate on more interesting tasks, where their specialist skills are really needed.
- 🗐 Speech-to-Speech MT is still a research topic. In general, there are many open research problems to be solved before MT systems will be come close to the abilities of human translators.
- ✵ Not only are there are many open research problems in MT, but building an MT system is an arduous and time consuming job, involving the construction of grammars and very large monolingual and bilingual dictionaries. There is no 'magic solution' to this.
- ▣ In practice, before an MT system becomes really useful, a user will typically have to invest a considerable amount of effort in customizing it.

✍ **6.** The philosopher Bar-Hillel in a ***1959*** report argued that fully automatic, high quality, MT (FAHQMT) was impossible, not just at present, but **in principle**. The problem he raised was that of finding the right translation for pen in a context like the following:

🖹 *Little John was looking for his toy*
box. Finally he found it. The box

*was in the **pen**. John was very happy.*

7. The argument was that (i) here **pen** could only have the interpretation play-pen, not the alternative writing instrument interpretation, (ii) this could be critical in deciding the correct translation for **pen**, (iii) discovering this depends on general knowledge about the world, and (iv) there could be no way of building such knowledge into a computer.

8. The doubts of funding authorities were voiced in the report which the US National Academy of Sciences commissioned in 1964 when it set up the Automatic Language Processing Advisory Committee (**ALPAC**) to report on the state of play with respect to MT as regards quality, cost, and prospects, as against the existing cost of, and need for translation. Its report, the so-called **ALPAC Report**, was damning, concluding that there was no shortage of human translators, and that there was no immediate prospect of MT producing useful translation of general scientific texts. This report led to the virtual end of Government funding in the USA. Worse, it led to a general loss of morale in the field, as early hopes were perceived to be groundless.

9. It was not until the late ***1970s*** that MT research underwent something of a **renaissance**. There were several signs of this renaissance. The Commission of the European Communities (CEC) purchased the English-French version of the SYSTRAN system, a greatly improved descendent of the earliest systems developed at Georgetown University (in Washington, DC), a Russian-English system whose development had continued throughout the lean years after ALPAC, and which had been used by both the USAF and NASA.

10. At about the same time, there was a rapid expansion of MT activity in ***Japan***, and the CEC also began to set up what was to become the EUROTRA project, building on the work of the GETA and SUSY groups. This was perhaps the largest, and certainly among the most ambitious research and development projects in Natural Language Processing.

11. In the ***late 1970s*** the Pan American Health Organization (PAHO) began development of a Spanish-English MT system (SPANAM), the United States *Air Force* funded work on the METAL system at the Linguistics Research Center, at the University of Texas in Austin, and the results of work at the TAUM group led to the installation of the METEO system. For the most part, the history of the ***1980s*** in MT is the history of these initiatives, and the exploitation of results in neighbouring disciplines.

12. Machine Translation and the Roller Coaster of History:

Chapter 2
Machine Translation in Practice

✍ **1.** A human translator will often be able to turn a badly written text into a well written translation; an MT system certainly will not. Bad input means bad output. Exactly what constitutes good input will vary a little from system to system. **Basic Writing Rules**:

- ☸ Keep sentences short.
- ⅄ Make sure sentences are grammatical.
- ❁ Avoid complicated grammatical constructions.
- ✠ Avoid (so far as possible) words which have several meanings.
- ⌘ In technical documents, only use technical words and terms which are well established, well defined and known to the system.

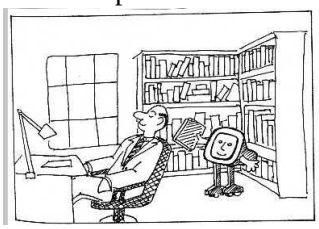
✍ **2.** In the past few years special tools have become available for supporting the production of text according to certain writing rules. There are spelling checkers and grammar checkers which can highlight words that are spelled incorrectly, or **grammatical** errors. There are also **critiquing systems** which analyse the text produced by an author and indicate where it deviates from the norms of the language. For example, given the example above of an over-complex sentence in a printer manual, such a tool might produce the following output:

Text Critique

New toner units are held level during installation and, since they do not as supplied contain toner, must be filled prior to installation from a toner cartridge.

Sentence too long.
during installation — disallowed use of word: installation.
prior — disallowed word.
since — disallowed clause in middle of sentence.

✍**3.** Translation Aids in the Workplace—***Automatic Lexical Lookup***:



Chapter 3
Representation and Processing

✍**1.** Human Translators actually deploy at least five distinct **kinds of knowledge**:

✪ Knowledge of the **source language**.

※ Knowledge of the **target language**. This allows them to produce texts that are acceptable in the target language.

◈ Knowledge of various **correspondences** between source language and target language (at the simplest level, this is knowledge of how individual words can be translated).

✳ Knowledge of the **subject matter**, including ordinary general knowledge and 'common sense'. This, along with knowledge of the source language, allows them to understand what the text to be translated means.

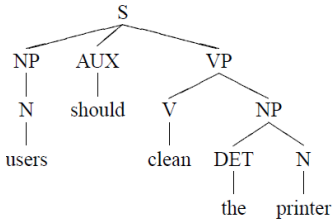
☆ Knowledge of the **culture**, **social** conventions, customs, and expectations, etc. of the speakers of the source and target languages.

✍**2.** In general, syntax is concerned with two slightly different sorts of **analysis** of sentences.

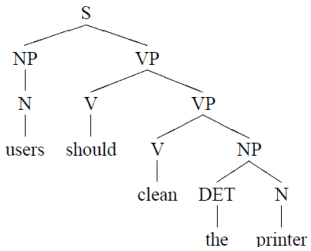
❄ The first is **constituent** or **phrase structure** analysis—the division of sentences into their constituent parts and the categorization of these parts as nominal, verbal, and so on.

⌖ The second is to do with **grammatical relations**; the assignment of grammatical relations such as SUBJECT, OBJECT, HEAD and so on to various parts of the sentence.

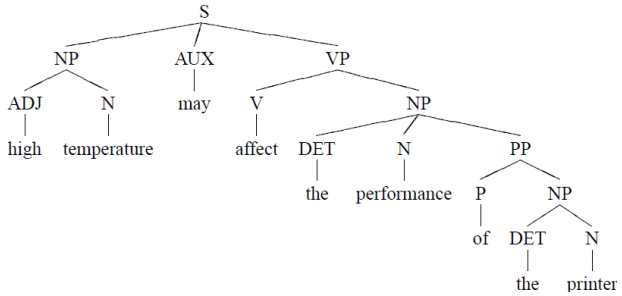
✍**3.** 'Sentence', is often abbreviated to S, 'noun phrase' to NP, 'verb phrase' to VP, 'auxiliary' to AUX, and 'determiner' to DET. This information is easily visualized by means of a labelled bracketing of a string of words, as follows, or as a **tree diagram**, as in *Figure 3.1*.



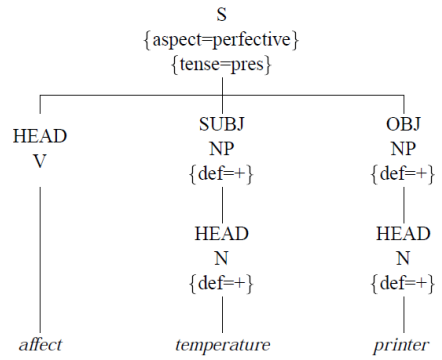
✍**4.** *Figure 3.2* **An Alternative Analysis**:



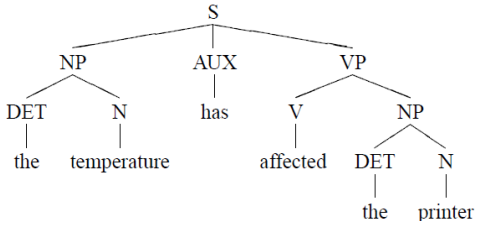
✍**5.** *Figure 3.3* A **More Complex** Tree Structure:



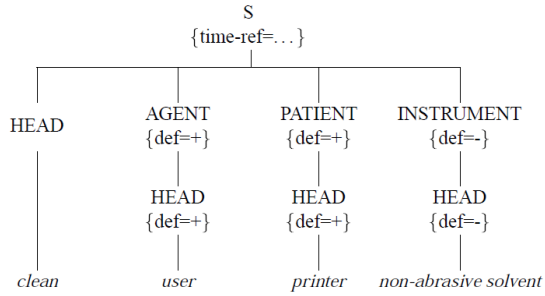
✍**6.** *Figure 3.4* A Representation of **Grammatical Relations** (The temperature has affected the printer):



✍**7.** *Figure 3.5* A **Constituent Structure** Representation:



✍**8.** *Figure 3.6* A Representation of **Semantic Relations** (The user cleans the printer with a non-abrasive solvent):



✍**9.** ***Knowledge*** can be ***manipulated automatically*** in two stages:

- 🗐 First, we will look at what is called **analysis**, or **parsing**. This is the process of taking an input string of expressions, and producing representations of the kind we have seen in the previous section.
- ✵ Second, we will look at **synthesis**, or **generation**, which is the reverse process—taking a representation, and producing the corresponding sentence.

✍**10.** The task of an **automatic parser** is to take a formal grammar and a sentence and apply the grammar to the sentence in order to

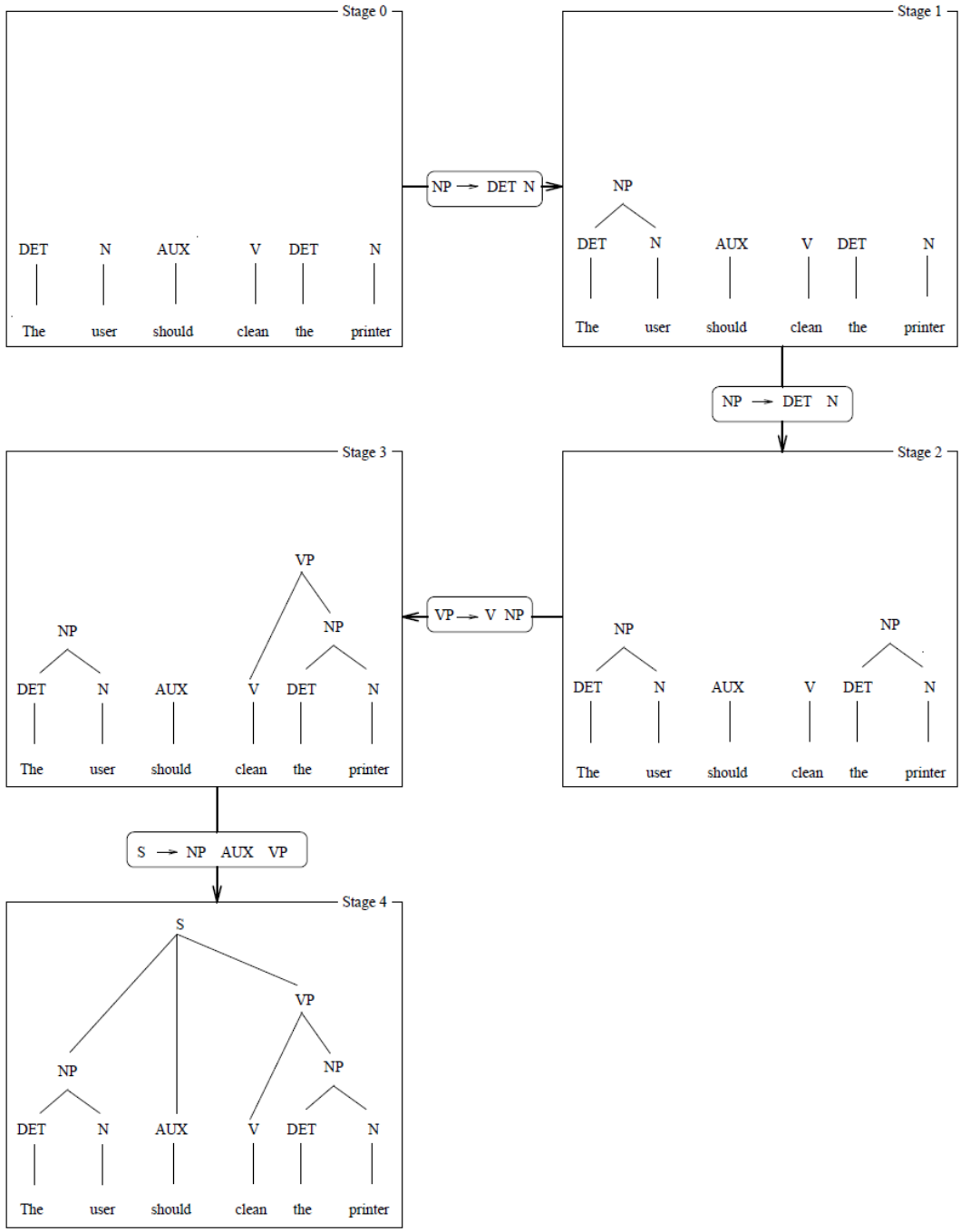
- ▣ check that it is indeed grammatical and

- given that it is grammatical, show how the words are combined into phrases and how the phrases are put together to form larger phrases (including sentences).

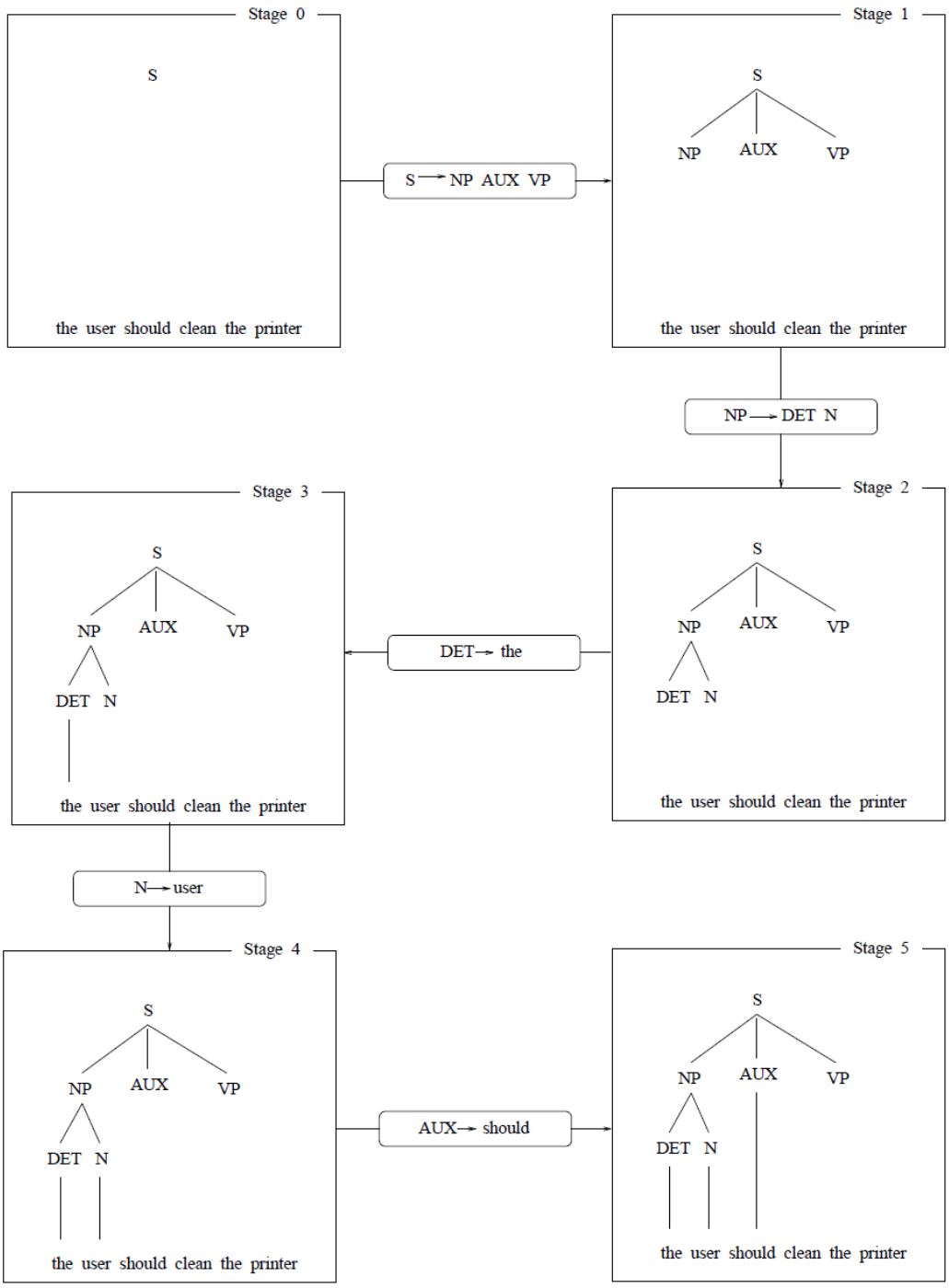
11. There are many ways to apply the **rules** to the input to produce an output tree—many different procedures, or parsing algorithms by which an input string can be assigned a structure. Here is one method:

1. For each word in the sentence, find a rule whose right hand side matches it. This means that every word would then be labelled with its part of speech (shown on the left hand side of the rule that matched it). This step is exactly equivalent to looking up the words in an English dictionary. Given rules of the type N=>user, N=>printer, and V=>clean, this will produce a partial structure as we can see at the top left corner (Stage 0) of *Figure 3.7*.
2. Starting from the left hand end of the sentence, find every rule whose right-hand side will match one or more of the parts of speech (Stage 1 of *Figure 3.7*).
3. Keep on doing step 2, matching larger and larger bits of phrase structure until no more rules can be applied. (In our example, this will be when the sentence rule finally matches up with a noun phrase and a verb phrase which have already been identified). The sentence is now parsed (Stage 2-4 of *Figure 3.7*).

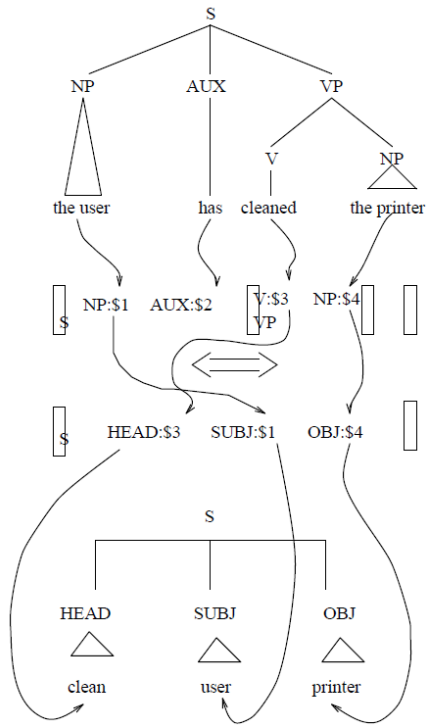
12. *Figure 3.7* Parsing Using a **Bottom-Up** Algorithm:



13. *Figure 3.8* Parsing Using a **Top-Down** Algorithm:


Stage 0
S
the user should clean the printer
S → NP AUX VP
Stage 1
S
NP AUX VP
the user should clean the printer
NP → DET N
Stage 2
S
NP AUX VP
DET N
the user should clean the printer
DET → the
Stage 3
S
NP AUX VP
DET N
the user should clean the printer
N → user
Stage 4
S
NP AUX VP
DET N
the user should clean the printer
AUX → should
Stage 5
S
NP AUX VP
DET N
the user should clean the printer

14. *Figure 3.9* Building a Representation of **Grammatical Relations**:

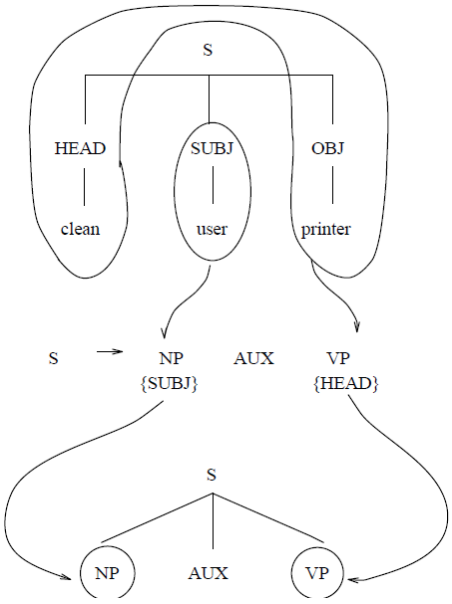


15. If the relations between syntactic, grammatical relation structures, and semantic structures are described by means of explicit rules, then one approach is to use those rules in the same way as we described for parsing, but 'in reverse'—that is with the part of the rule written after the "<=>" interpreted as the lhs. Things are not quite so straightforward when information about grammatical relations, and/or semantics is packed into the constituent structure rules.

16. One possibility is to have a completely separate set of procedures for producing sentences from semantic or grammatical relation structures, without going through the constituent structure stage (for example, one would need a rule that puts **HEAD**, **SUBJECT**, and **OBJECT** into the normal word order for English, depending on whether the sentence was active or passive, interrogative or declarative). This has attractions, in particular, it may be that one does not want to be able to generate exactly the sentences one can parse (one may want one's parser to accept stylistically rather bad sentences, which one would not want to produce, for example). However, the disadvantage is that one will end up describing again most, if not all, of the knowledge that is contained in the grammar which is used for parsing.

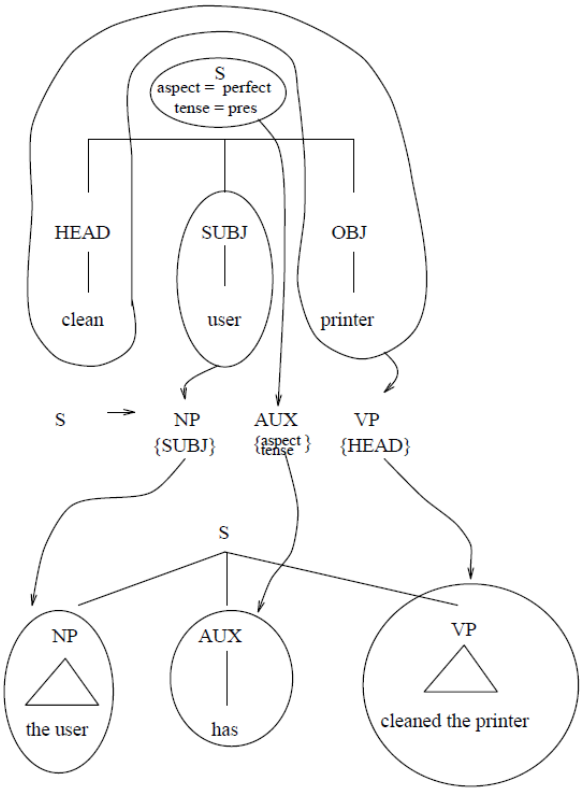
17. A naive (and utterly impractical) approach would be to simply apply **constituent structure** rules at random, until a structure was produced that matched the grammatical relation structure that is input to generation. A useful variation of this is to start with the whole input structure, and take all the rules for the category **S** (assuming one expects the structure to represent a sentence), and to compare the grammatical relation structure each of these rules produces with the input structure. If the structure produced by a particular rule matches the input structure, then build a partial tree with this rule, and mark each of these parts as belonging to that tree. For example, given the rule for **S** above, one could take the grammatical relation structure of a sentence like The user has cleaned the printer and begin to make a phrase structure tree, as is illustrated in *Figure 3.10*.

18. *Figure 3.10* Generation from a **Grammatical Relation** Structure 1:



19. One can see that a partial constituent structure tree has been created, whose nodes are linked to parts of the grammatical relation structure (a convention is assumed here whereby everything not explicitly mentioned in the rule is associated with the **HEAD** element). Now all that is necessary is to do the same thing to all the parts of the Grammatical relation structure, attaching the partial trees that have been constructed in the appropriate places. This is illustrated in *Figure 3.11*. Again, there are many refinements and details missed out here, but again, all that matters is the basic picture.

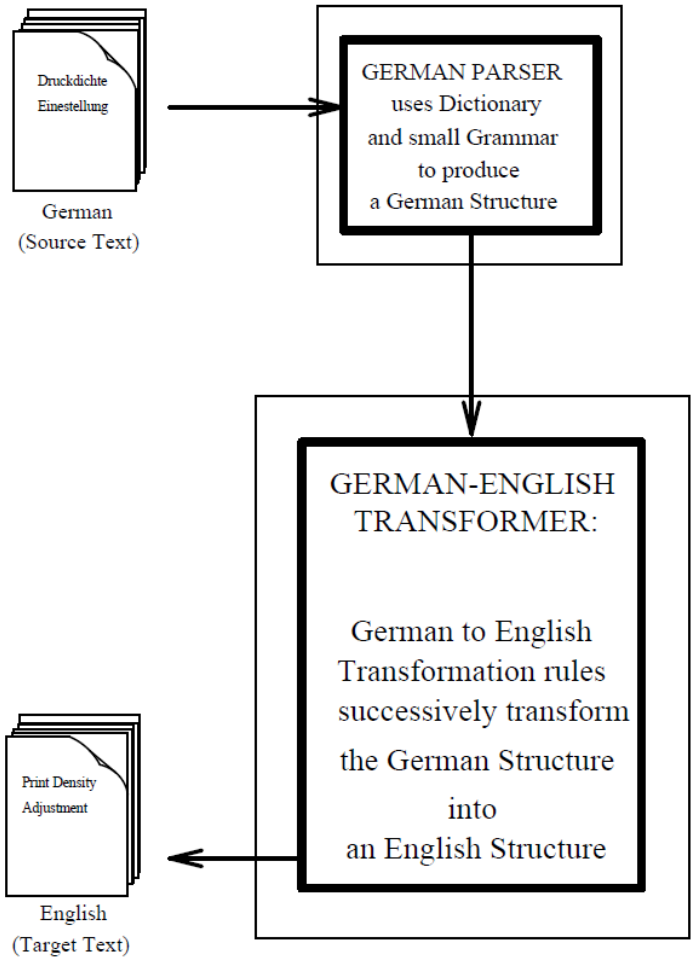
✍**20.** *Figure 3.11* Generation from a Grammatical Relation Structure 2:



Chapter 4
Machine Translation Engines

✍**1.** The main idea behind **transformer engines** is that input (source language) sentences can be transformed into output (target language) sentences by carrying out the simplest possible parse, replacing source words with their target language equivalents as specified in a bilingual dictionary, and then roughly re-arranging their order to suit the rules of the target language. The overall arrangement of such an Engine is shown in *Figure 4.1*.

✍**2.** *Figure 4.1* A **Transformer Architecture** (German to English):



✍ **3.** We can summarise the situation of the **transformer engine architecture** as follows:

- 🗎 It is highly **robust**. That is, the Engine does not break down or stop in an 'error condition' when it encounters input which contains unknown words or unknown grammatical constructions. Robustness is clearly important for general-purpose MT.
- ☸ In the worst case it can work rather badly, being prone to produce output that is simply unacceptable in the target language ('word salad').
- ✠ The translation process involves many different rules interacting in many different ways. This makes transformer systems rather hard to understand in practice—which means that they can be hard to extend or modify.

- The transformer approach is really designed with translation in one direction, between one pair of languages in mind, it is not conducive to the development of genuinely multi-lingual systems (as opposed to mere collections of independent onepair, one-direction engines).

4. The second major architecture—**indirect** or **linguistic knowledge** (LK) architecture—has dominated research in MT design during the past decade and is starting to appear in a number of commercial systems. The idea behind LK engines is straightforward enough: High quality MT requires linguistic knowledge of both the source and the target languages as well as the differences between them.

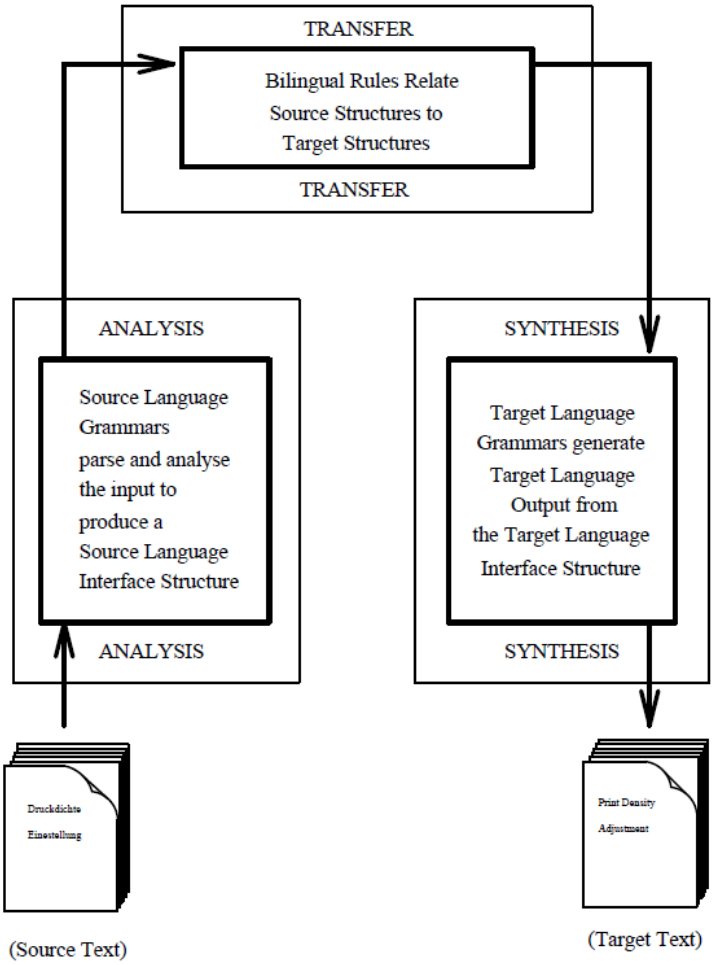
5. With the **Transformer architecture**, the translation process relies on some knowledge of the source language and some knowledge about how to transform partly analysed source sentences into strings that look like target language sentences.

6. With the **LK architecture**, on the other hand, translation relies on extensive **knowledge** of both the source and the target languages and of the relationships between analysed sentences in both languages.

7. LK architecture typically accords the target language the same status as the source language. As can be seen from *Figure 4.2*, the **LK architecture** requires two things:

- A substantial grammar of both the source language and the target language. These grammars are used by parsers to analyse sentences in each language into representations which show their underlying structure, and by generators to produce output sentences from such representations.
- An additional comparative grammar which is used to relate every source sentence representation to some corresponding target language representation—a representation which will form the basis for generating a target language translation.

8. *Figure 4.2* The Components of a **Transfer System**:



9. Looking at *Figure 4.2*, it is clear that if (say) the system is translating from German to English:

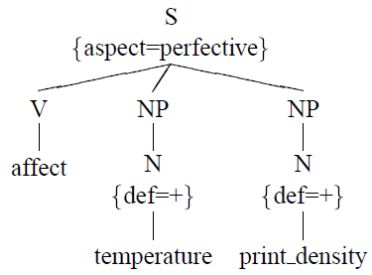
- ※ The **first** (*analysis*) step involves using the parser and the German grammar to analyse the German input.
- ◈ The **second** (*transfer*) step involves changing the underlying representation of the German sentence into an underlying representation of an English sentence.
- ✳ The **third** (*synthesis*) step and final major step involves changing the underlying English representation into an English sentence, using a generator and the English grammar.

10. The fact that a proper English **grammar** is being used means that the **output** of the system—the English sentences—are far more likely to be *grammatically correct* than those of a **German-English Transformer**

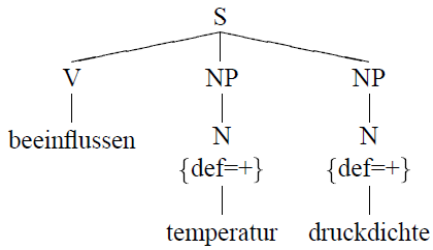
system (recall that the latter had no explicit English grammar to guide it). In fact, if (*per impossibile*) we had an **LK German-English** system with a 'perfect' English grammar the only sort of mistake it could make in the output would be errors in translational accuracy. That is, it would always produce perfectly well-formed English sentences even when it did not produce correct translations.

✍**11.** If we are translating the sentence "*The temperature has affected the print density*" into German, the **analysis** component might produce a representation along the lines of *Figure 4.3*.

✍**12.** *Figure 4.3* Abstract Tree Representation:



✍**13.** *Figure 4.4* Tree Representation after Translation: [Thus, "*Die Temperatur hat die Druckdichte beeinflußt*" should be produced as the translation.]



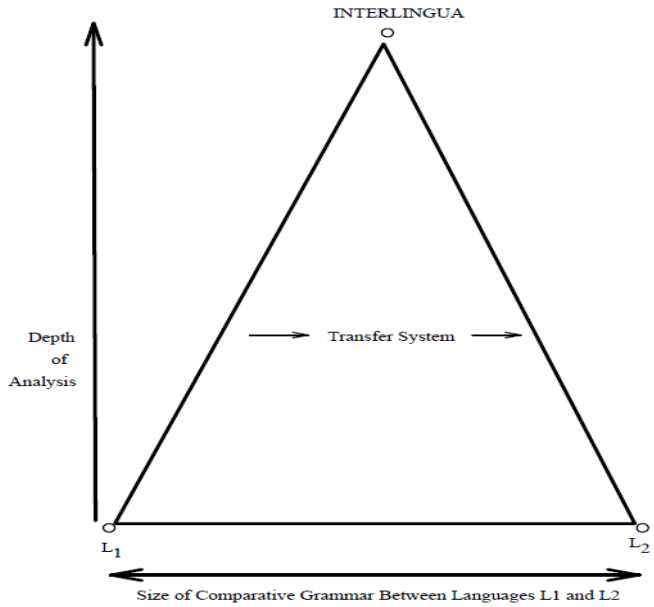
✍**14.** A **Transformer engine** generally preserves the surface order of the source language and directly re-uses it—with modifications where appropriate—to order the target language words. An **LK engine**, on the other hand, extracts all the information it can from the source word order and recodes this information in a more or less abstract representation.

✍**15.** The generator for the target language will use the information in the representation and in the target language grammar to construct a target language sentence with a word order that it is grammatically appropriate for that language. In short, ordering information is **not** normally carried over **directly**.

✍**16.** A major objective of **MT** research is to define a level of **analysis** which is so deep that the comparative grammar component disappears completely. Given such a level of representation, the output of **analysis**

could be the direct input to the target **synthesis** component. Representations at such a level would have to capture whatever is common between sentences (*and expressions of other categories*) and their translations—that is they would have to be representations of 'meaning' (in some sense). Moreover, such a level of representation would have to be entirely **language independent**—for example, if it preserved features of the source language, one would still require a transfer component of some kind to produce the corresponding features of the target language. For this reason, such a level of representation is normally called an **Interlingua**, and systems that use such a level are called **Interlingual**.

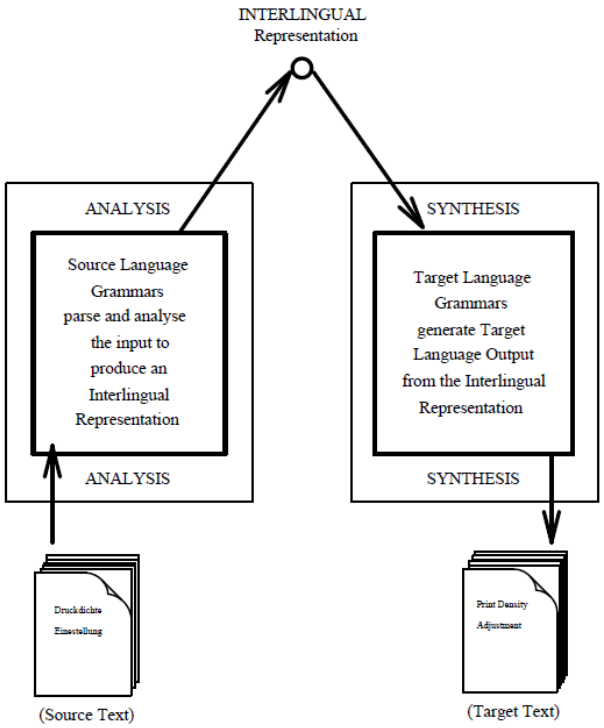
✍**17.** The relationship between ***transfer*** and ***interlingual*** systems can be pictured as in Figure 4.6. As one can see, the size of the **contrastive grammar** (hence the transfer component) between two languages decreases as the level of representation becomes more abstract. As this diagram perhaps suggests, the difference between **transfer** representations and **interlinguas** is a matter of degree rather than absolute distinction:



✍**18.** The size of the **comparative grammar** that is required to translate between two languages gets smaller as the 'depth' of the representations used increases. As the representations become more abstract, there are fewer differences between source and target representations and it is easier to relate them. Ultimately, a level of representation may be achieved where source and target representations are identical, where **no**

comparative grammar is needed. In this situation, the representations which are produced by analysis could be directly input to the target language synthesis component. Such a level of representation is called an **interlingua**, and a system that uses such a level is called an **interlingual system**.

✍ **19.** *Figure 4.7* The Components of an **Interlingual System**:



✍ **20.** The performance characteristics of an **LK engine**:

- ✯ Because the system has a (*partial*) **grammar** of the target language, output will tend to be grammatical. At any rate, it will be far less strange and far less source-language grammar- dependent than output from transformer engines.
- ❄ Because the **comparative grammar** completely specifies a relationship between representations of two languages, translational quality will tend to be more reliable than for transformer engines.
- ⌖ Because the system tends to separate language into separate modules (one grammar for each language and one comparative grammar for each pair of languages), it is relatively easy in principle to add new languages to the system. For example, adding Dutch to a German-English system would require only the addition of a Dutch grammar module and Dutch-English and German-English comparative

grammar modules. Individual language modules can be designed and constructed without specifying which other language modules they will have to work with in the final system. Of course, this matters more to the developer than the user since it is the former that writes and supplies basic language modules.

- The system will be upset by unusual, marginally acceptable or frankly unacceptable input sentences because it has a grammar for the source language and hence a strong notion of grammaticality.
- Because the grammars that computational linguists are able to write are invariably less complete than the 'real' complete grammar of any language, there will be some complicated grammatical input sentences that the system fails to recognise.

21. From the engine manufacturer's point of view, the **transformer architecture** has the advantage that it accepts anything that is given to it (though the translations it produces are another matter). The **LK architecture** is at a disadvantage here: because it thinks it knows something about the languages involved, it tends to think that anything it doesn't know isn't language and hence unacceptable. As a consequence, a pure **LK engine** during its development phase tends to grind to a halt on anything unusual, or even on something quite common which the developer has forgotten to include.

Chapter 5
Dictionaries

1. Dictionaries are the largest components of an MT system in terms of the amount of information they hold. If they are more then simple word lists (and they should be, if a system is to perform well), then they may well be the most expensive components to construct.

2. More than any other component, the **size** and **quality** of the ***dictionary*** limits the scope and coverage of a system, and the quality of translation that can be expected.

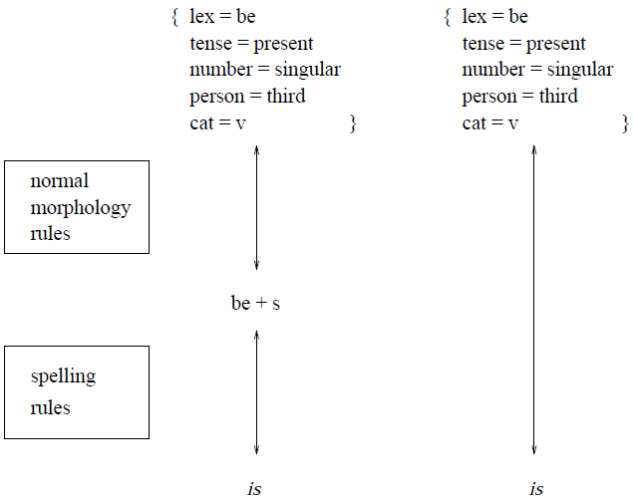
3. The **dictionaries** are where the end user can expect to be able to contribute most to a system—in fact, an ***end user*** can expect to have to make some additions to system dictionaries to make a system really useful. While MT suppliers rarely make it possible for users to modify other components, they normally expect them to make additions to the dictionary. Thus, from the point of view of a user, a basic understanding of dictionary construction and sensitivity to the issues involved in 'describing words' is an important asset.

4. Subcategorization information indicates that, for example, the verb button occurs with a noun phrase OBJECT. In fact, we know much more about the verb than this—the OBJECT, or in terms of semantic roles, the PATIENT, of the verb has to be a 'buttonable' thing, such as a piece of clothing, and that the SUBJECT (more precisely AGENT) of the verb is normally animate.5 Such information is commonly referred to as the **selectional restrictions** that words place on items that appear in constructions where they are the HEAD.

5. Morphology is concerned with the internal structure of words, and how words can be formed. It is usual to recognize three different word formation processes:

- **Inflectional processes**, by means of which a word is derived from another word form, acquiring certain grammatical features but maintaining the same part of speech or category (e.g. walk, walks);
- **Derivational processes** in which a word of a different category is derived from another word or word stem by the application of some process (e.g. grammar=>grammatical, grammatical=>grammaticality);
- **Compounding**, in which independent words come together in some way to form a new unit (buttonhole).

6. *Figure 5.2* Treatment of Irregular Verbs:



Chapter 6
Translation Problems

1. We consider problems under the following headings:

- Problems of **ambiguity**,

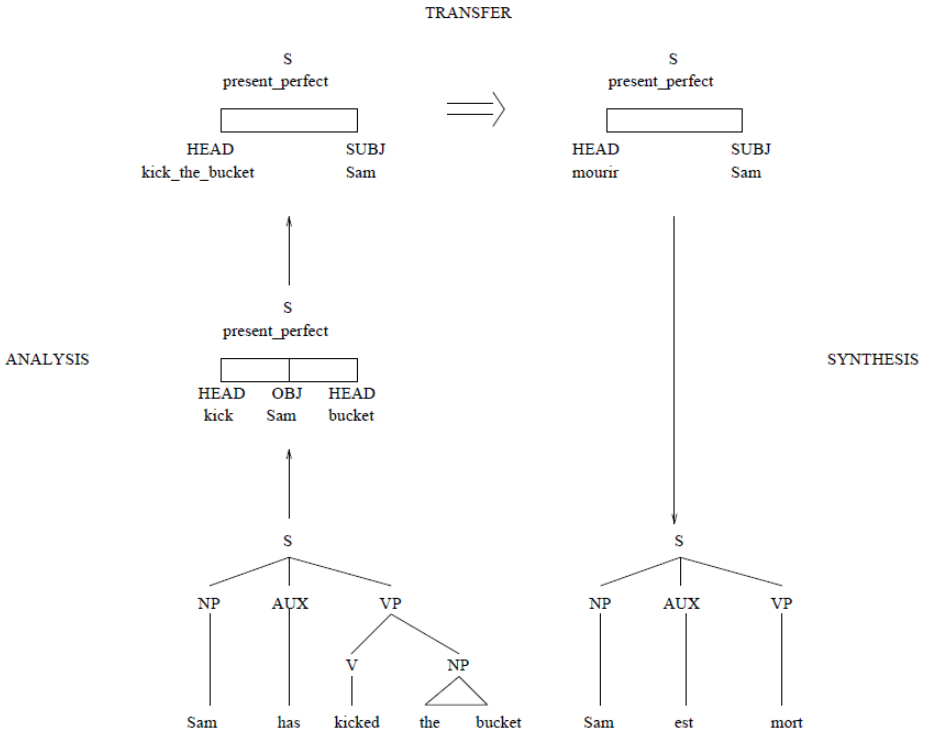
- problems that arise from **structural** and **lexical** differences between languages and
- **multiword units** like idioms and collocations.

2. Of course, these sorts of problem are not the only reasons why MT is hard. Other problems include the sheer **size** of the undertaking, as indicated by the number of **rules** and **dictionary entries** that a realistic system will need, and the fact that there are many constructions whose **grammar** is poorly understood, in the sense that it is not clear how they should be represented, or what rules should be used to describe them.

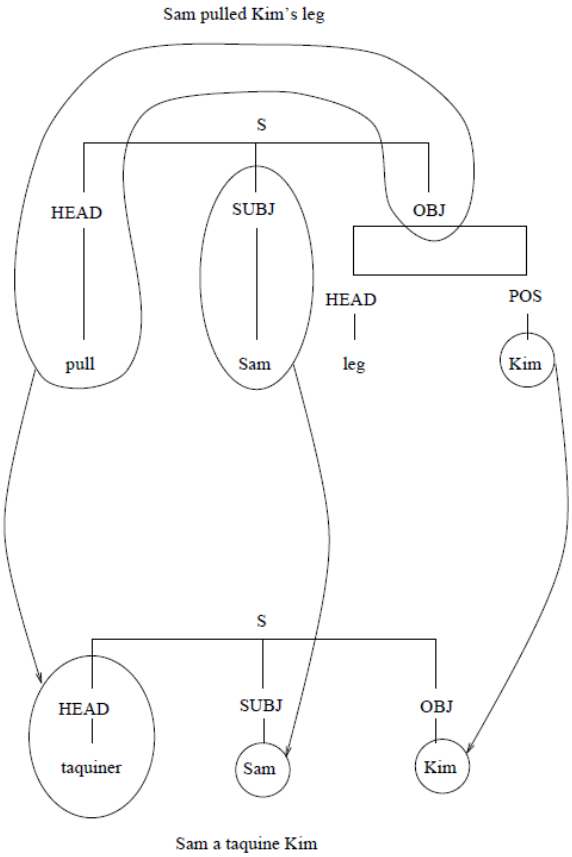
3. When a word has more than one meaning, it is said to be **lexically ambiguous**. When a phrase or sentence can have more than one structure it is said to be **structurally ambiguous**.

4. Lexical holes refer to cases where one language has to use a phrase to express what another language expresses in a single word. Examples of this include the '*hole*' that exists in English with respect to French ignorer ('to not know', 'to be ignorant of'), and se suicider ('to suicide', i.e. 'to commit suicide', 'to kill oneself'). The problems raised by such **lexical holes** have a certain similarity to those raised by idioms: in both cases, one has phrases translating as single words.

5. *Figure 6.4* Dealing with **Idioms ❶**:



✍**6.** *Figure 6.5* Dealing with **Idioms ❷**:



Chapter 7

Representation and Processing Revisited: Meaning

✍ **1.** It is useful to think of the kind of **knowledge** that systems are equipped with as being of three kinds:

- ✠ linguistic knowledge which is independent of context, **semantic knowledge**
- ⌘ linguistic knowledge which relates to the context (e.g. of earlier utterances), sometimes called **pragmatic knowledge**
- ✪ common sense, general, non-linguistic knowledge about the real world, which we will call real **world knowledge**

✍ **2.** There are many ways of thinking about and representing word meanings, but one that has proved useful in the field of machine translation involves associating words with **semantic features** which correspond to their sense components:

```
man = (+HUMAN, +MASCULINE and +ADULT)

woman = (+HUMAN, -MASCULINE and +ADULT)

boy = (+HUMAN, +MASCULINE and -ADULT)

girl = (+HUMAN, -MASCULINE and -ADULT)
```

✍ **3.** Associating words with semantic features is useful because some words impose **semantic constraints** on what other kinds of words they can occur with. For example, the verb eat demands that its AGENT (the eater) is animate and that its PATIENT (that which is eaten) is edible, — concrete (rather than abstract, like sincerity, or beauty), and solid (rather than liquid, so one cannot 'eat' beer, coffee, etc.; soup is a borderline case).

✍ **4.** We can ***encode*** this constraint in our grammar by associating the features HUMAN and EDIBLE with appropriate nouns in our dictionary and describing our entry for eat as something like cat=verb, AGENT=HUMAN, PATIENT=EDIBLE. The grammar will now only accept objects of eat that have the feature EDIBLE. Thus these **selectional restrictions**, as they are called, act as a filter on our grammar to rule out unwanted analyses.

✍ **5. Anaphoric pronouns** are those which refer back to some **antecedent** earlier in the text, as the pronoun it in (❶) refers back to its antecedent the cake. [❶ *Sam took the cake from the table. Then he ate it.*] Take the translation of (❶) from English into French. We know that it must refer back to some singular noun in the previous text or discourse. It has been shown that it is very often the case that the antecedent of a pronoun is in the same sentence or in the immediately preceding sentence. Assuming that these are the first sentences in our text, then it can potentially refer back to one of three NPs, namely ***Sam***, ***the cake*** or ***the table***. The syntactic facts of English constrain the pronoun to agree in number and gender with its antecedent, so it being a neuter pronoun cannot possibly refer to ***Sam***, which is either masculine or feminine. That leaves us with the choice of either cake or table.

✍ **6.** One might wonder at this stage whether we need to decide between the two at all, or whether we can preserve the ambiguity of it in translation. It turns out that French, like English, requires a pronoun to agree in number and gender with its antecedent. However, since ***cake*** translates as the masculine noun *gˆateau* in French and table as the feminine noun table, this means that we do have to decide which noun the pronoun it refers back to, in order to translate it either as le (where it would be interpreted as referring to *le gˆateau*—***cake***) or as la (where it would refer back to la table in the translation of the first sentence). In the above example we can

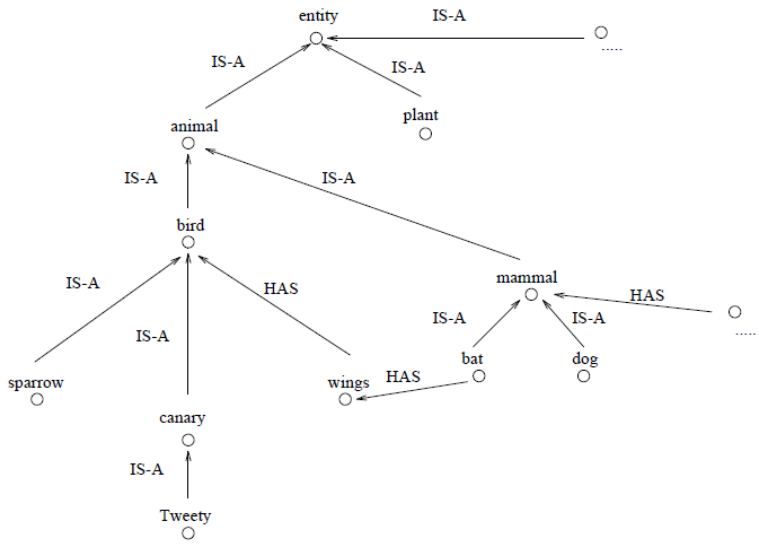
use **selectional restrictions** on the type of object that eat can have (namely 'edible' objects) to exclude, or at least '*disprefer*', ***table*** as an antecedent for it. This leaves ***cake*** as the best candidate. Providing rules which allow this sort of process to be performed automatically is not too difficult, but unfortunately resolving pronoun reference is not generally that simple.

✍ **7.** Faced with two competing candidates for *pronominal reference* in a segment, there is another fact about **discourse** that we can exploit to get at their resolution, and this is the notion of **focus**. At any time in a discourse segment there is an object which is the prime candidate for pronominal reference, and this element is called the focus. Different suggestions have been made as to how to identify the focus. Often, there are syntactic signals. For example, in the following example, the focus is much more likely to be Kim, than Sam, and Kim is more likely to be the antecedent of a pronoun in the following sentence. [❷ *It was Kim who Sam telephoned.* ***She*** *was in the bath.*] The **focus** of a sentence is also often the NP that has the THEME role in the previous sentence (the THEME role includes what we have been calling the PATIENT role, but is slightly more general). This is the case with Kim in (❷), which reinforces the structural cue.

✍ **8.** But even in the following sequence, where there are no clear ***structural clues***, key is the THEME and hence most likely to be the focus of the first sentence (and therefore key is preferred to doormat as the referent of it in the second sentence). [❸ *She put the key under the doormat. When she came home, she found that* ***it*** *had been stolen.*]

✍ **9.** Unlike most *linguistic* knowledge, in particular, most knowledge of syntax and semantics, **real world knowledge** is generally '**defeasible**', that is, subject to revision, and not guaranteed correct – humans have little trouble assuming one thing most of the time, but managing with a contradictory assumption on occasions. This is extremely difficult to automate. A second problem is the huge amount of such knowledge we seem to have (knowledge about the relative sizes of almost everything, for example).

✍ **10.** However, there are some **methods** of representation that are useful for some kinds of knowledge. One particularly useful representation is the so called **Semantic Net** which can be used for representing 'is a' relations (such as 'a dog is a mammal'). *Figure 7.4* gives a small part of such a network:



✍ **11.** Intuitively, the nodes in such a **network** stand for things, and the links between them are **relations**. This means that it can easily be generalized for other sorts of relations. For example, adding other objects, and using a 'part of' relation, one could represent the fact that (say) a printer is made up of various components, and the fact that these are in turn made up of subcomponents, etc. Such information might be important in understanding sentences like the following: [▣ *Put the toner in the cartridge in the reservoir.*] Knowing that the *reservoir* does not have a *cartridge* as a part would allow one to work out that this is an instruction to put the toner which is in the *cartridge* in the *reservoir*, rather than an instruction to put the toner in a particular *cartridge* (i.e. the one that is in the reservoir).

Chapter 8
Input

✍ **1.** In this chapter, we describe how the full potential of **machine readable texts** can be exploited in three ways:

- ※ first, by adopting the notion of an 'electronic document' and embedding an MT system in a complete document processing system
- ◈ second, by restricting the form of input by using simplified or **controlled language**
- ✳ third, by restricting both the form, and the subject matter of the input texts to those that fall within a **sublanguage**—it is here that the immediate prospects for MT are greatest

✍ **2.** The common theme of this chapter is ***how*** the successful application of MT can be enhanced by ensuring that the input to the system is '**appropriate**'. Briefly, the message is this: having texts in machine readable form is a prerequisite for sensible use of MT, but one can get much better results by:

- ✰ adopting certain standard formats for the input
- ❄ controlling the input, so that problematic constructions, etc., are avoided
- ⌖ tailoring the MT systems to the language of particular domains

✍ **3.** The reasons for **controlled languages**' superior MT performance are easy to understand.

- 🗐 First, the restricted vocabulary means that fewer words need to be added to the MT system dictionaries and more effort can be put into getting the entries which are required right.
- ✵ Second, the grammar component of the system can be tailored to handle all and only those constructions which are licensed by the controlled language specification, a specification which excludes the most difficult and ambiguous constructions anyway.

✍ **4.** The PACE (Perkins Approved Clear English) Writing Rules:

✺ **Keep it short and simple:**

1 Keep sentences short.
2 Omit redundant words.
3 Order the parts of the sentence logically.
4 Don't change constructions in mid-sentence.
5 Take care with the logic of *and* and *or*.

❁ **Make it explicit:**

6 Avoid elliptical constructions.
7 Don't omit conjunctions or relatives.
8 Adhere to the PACE dictionary.
9 Avoid strings of nouns.
10 Do not use *-ing* unless the word appears thus in the PACE dictionary.

✍ **5.** A sample from the PACE Dictionary:

advantage	n	Benefit
adverse	adj	Unfavourable
advice	n	Specialist Intelligence
advise,d	v	To provide advice
aerosol container	n	
affect,ed	v	To have an effect on
after	adv,prep	Being behind in succession, following something
again	adv	Once more
against	prep	In contact with
agglomerator	n	
agricultural	adj	Appertaining to agriculture
air	n	The gases that surround the earth
air charge cooler	n	

6. The Effect of Using Controlled English:

- **BEFORE:** It is equally important that there should be no seasonal changes in the procedures, as, although aircraft fuel system icing due to water contaminations more often met with in winter, it can be equally dangerous during the summer months.
- **AFTER:** Use the same procedure all the time, because water in the fuel system can freeze during winter or summer.
- **BEFORE:** Loosen the dynamo or alternator mounting and adjustment link fasteners.
- **AFTER:** Loosen the pivot fasteners of the dynamo or alternator mounting. Loosen also the fasteners of the adjustment link.
- **BEFORE:** Reference to renewing the joints and cleaning of joint faces has to a great extent been omitted from the text, it being understood that this will be carried out where applicable.
- **AFTER:** Normally the text does not include instructions to clean joint faces or to renew joints. These operations must be done, if necessary.

7. The term **sublanguage** refers to the **specialized language** used (predominantly for communication between experts) in certain fields of knowledge, for example, the language of weather reports, stockmarket reports, the language of some kinds of medical discussion, the language of aeronautical engineering. **Specialized vocabulary** is one characteristic of such 'languages' (they typically contain words not known to the non-specialist and also words used in different or more precise ways).

8. However **sublanguages** are also often characterised by special or restricted grammatical patterns. In MT, it is quite common to use the term **sublanguage** rather loosely to refer not just to such a **specialized language**, but to its use in a particular type of text (e.g. installation manuals, instruction booklets, diagnostic reports, learned articles), or

with a particular communicative purpose (communication between experts, giving instructions to non-experts, etc).

Chapter 9
Evaluating MT Systems

✍ **1.** A traditional way of assessing the quality of translation is to assign scores to output sentences. A common aspect to score for is **Intelligibility**, where the intelligibility of a translated sentence is affected by grammatical errors, mistranslations and untranslated words.

✍ **2.** The major MT evaluation studies which have been published report on different scoring systems; the number of points on the scoring **scales** ranging from 2 (intelligible, unintelligible) to 9. The 9 point scale featured in the famous ***ALPAC Report*** and was not just used to score the **intelligibility** of MT, but also of human translation. As a consequence the **scale** included judgments on fairly subtle differences in e.g. style. This **scale** is relatively well-defined and well-tested. Nevertheless we think that it is too fine-grained for MT evaluation and leads to an undesirable dispersion of scoring results. Also, we think that **style** should not be included because it does not affect the **intelligibility** of a text. On the other hand, a two point scale does not give us enough information on the seriousness of those errors which affect the **intelligibility**. (A two point scale would not allow a distinction to be drawn between the examples in the previous paragraph, and complete garbage, (or something completely untranslated) and a fully correct translation.) Perhaps a ***four point scale*** like the one below would be more appropriate.

✍ **3. An *Example* Intelligibility Scale:**

⌘ The sentence is perfectly clear and intelligible. It is grammatical and reads like ordinary text.

✪ The sentence is generally clear and intelligible. Despite some inaccuracies or infelicities of the sentence, one can understand (almost) immediately what it means.

※ The general idea of the sentence is intelligible only after considerable study. The sentence contains grammatical errors and/or poor word choices.

☆ The sentence is unintelligible. Studying the meaning of the sentence is hopeless; even allowing for context, one feels that guessing would be too unreliable.

✍ **4.** By measuring intelligibility we get only a partial view of translation quality. A highly intelligible output sentence need not be a correct translation of the source sentence. It is important to check whether the meaning of the source language sentence is preserved in the translation.

This property is called **Accuracy** or **Fidelity**. Scoring for accuracy is normally done in combination with (but after) scoring for intelligibility.

✍**5.** In general it seems an **operational evaluation** conducted by a user will be extremely **expensive**, requiring 12 person-months or more of translator time. An attractive approach is to integrate the **evaluation process** in the normal ***production process***, the only difference being that records are kept on the number of input words, the turnaround time and the costs in terms of **time** spent in post-editing. The **cost** of such an integrated operational evaluation is obviously less. After all, if the system is really good the translation costs will have been reduced and will compensate for some of the costs of the evaluation method.

Chapter 10
New Directions in MT

✍**1.** Most transfer or interlingual **rule-based** systems are based on the idea that success in practical MT involves defining a level of representations for texts which is abstract enough to make translation itself straightforward, but which is at the same time superficial enough to permit sentences in the various source and target languages to be successfully mapped into that level of representation. That is, successful MT involves a compromise between depth of analysis or understanding of the source text, and the need to actually compute the abstract representation. In this sense, **transfer systems** are less ambitious than **interlingual systems**, because they accept the need for (often quite complex) mapping rules between the most abstract representations of source and target sentences. As our **linguistic knowledge** increases, so too MT systems based on **linguistic rules** encoding that knowledge should improve. This position is based on the fundamental assumption that finding a sufficiently abstract level of representation for MT is an attainable goal. However, some researchers have suggested that it is **not** always the case that the deepest level of representation is necessarily the best level for translation.

✍**2.** The term **knowledge-based MT** has come to describe a **rule-based system** displaying extensive semantic and pragmatic knowledge of a domain, including an ability to reason, to some limited extent, about concepts in the domain (the components, installation and operation of a particular brand of laser printer could constitute a domain). We noted the appeal of such an approach as a way of solving some basic MT problems in earlier chapters. Essentially, the premise is that **high quality** translation requires in-depth understanding of the text, and the development of the domain model would seem to be necessary to that

sort of deep understanding. One of the important considerations driving this work is an appreciation that **post-editing** is time-consuming and very expensive, and therefore that efforts made to produce high quality output will pay off in the long run.

3. One of the most serious problems, and probably the most serious problem, for **linguistic knowledge MT** is the development of *appropriate large-scale* grammatical and lexical resources. There are really a number of closely related problems here.

- The first is simply the **scale** of the undertaking, in terms of numbers of linguistic rules and lexical entries needed for fully automatic, high quality MT for **general** purpose and **specialised** language usage.
- The second concerns the difficulties of **manipulating** and **managing** such knowledge within a working system. The experience of linguists developing a wide variety of natural language processing systems shows that it is all too easy to add ad hoc, specially crafted rules to deal with problem cases, with the result that the system soon becomes difficult to understand, upgrade and maintain.
- The third issue is one of **quality** and concerns the level of linguistic detail required to make the various discriminations which are necessary to ensure high quality output, at least for general texts.

4. In developing MT systems, **bilingual texts** are an extremely important resource, and they are most useful if organized in such a way that the user can view translation 'chunks' or 'units'. In **bitext** (or '**multitext**') the text is aligned so that within each bilingual (or multilingual) chunk the texts are translations of each other. The most common form of alignment takes the sentence to be the organizing unit for chunking and techniques exist for performing this alignment of bitext automatically with a high level of accuracy (96% or higher). Of course alignment does not need to stop at the sentence level and it is possible to apply simple probability measures to a sentence aligned **bitext** to extract automatically the most probable word pair alignments, and given some skeleton or phrasal parsing, to attempt to extract useful information about phrasal alignment.

5. Throughout most of this book, we have assumed a model of the translation machine which involves explicit *mapping rules* of various sorts. In the '**translation by analogy**', or '**example-based**' approach, such *mapping rules* are dispensed with in favour of a procedure which involves matching against stored example translations. The basic idea is to collect a bilingual corpus of translation pairs and then use a best match algorithm to find the closest example to the source phrase in question. This gives a translation template, which can then be filled in by word-for-word translation.

✍**6.** A pure **example-based approach** would use no grammar rules at all, only example phrases. However, one could also imagine a role for some normal *linguistic analysis*, producing a standard linguistic representation. If, instead of being given in simple 'string' form, examples were stated in terms of such representations (i.e. given as fragments of linguistic representations), one would expect to be able to deal with many more variations in sentence pattern, and allow for a certain amount of restructuring in generation. In this way, one would have something that looked more like a standard **LK architecture**. The chief difference would be in the level of specificity of the rules. In particular, where in a traditional **transfer system** the rules are stated in as general a form as possible, to cover entire classes of case, what one would have here is a system where the rules are stated in highly particular forms (each one for essentially one case), but there is a general procedure for estimating, for each case, which rule is most appropriate (i.e. by estimating which example is closest). Of course, what this suggests is that there is no radical incompatibility between **example-based**, and **rule-based approaches**, so that the real challenge lies in finding the best combination of techniques from each.

✍**7.** With respect to MT, the term '**statistical approaches**' can be understood in a narrow sense to refer to approaches which try to do away with explicitly formulating linguistic knowledge, or in a broad sense to denote the application of statistically or probablistically based techniques to parts of the MT task (e.g. as a word sense disambiguation component).

✍**8.** The approach can be thought of as trying to apply to MT techniques which have been highly successful in ***Speech Recognition***, and though the details require a reasonable amount of statistical sophistication, the basic idea can be grasped quite simply. The two key notions involved are those of the **language model** and the **translation model**. The language model provides us with probabilities for strings of words (in fact sentences), which we can denote by **Pr (S)** (for a source sentence **S**) and **Pr (T)** (for any given target sentence **T**). Intuitively, **Pr (S)** is the probability of a string of source words **S** occurring, and likewise for **Pr (T)**.

✍**9.** The **translation model** also provides us with probabilities—**Pr (T|S)** is the conditional probability that target sentence |will occur in a target text which translates a text containing the source sentence **S**. The product of this and the probability of S itself, that is **Pr (S)×Pr (T|S)** gives the probability of source-target pairs of sentences occurring, written **Pr (S,T)**. One task, then, is to find out the probability of a source string (or sentence) occurring (i.e. **Pr (S)**). This can be decomposed into the

probability of the first word, multiplied by the conditional probabilities of the succeeding words, as follows:

$$\Pr(s1) \times \Pr(s2|s1) \times \Pr(s3|s1.s2), \text{ etc...}$$

$$\Pr(S|T) = \frac{\Pr(S)\Pr(T|S)}{\Pr(T)}$$

✍**10.** In order to get some idea of how the **translation model** works, it is useful to introduce some further notions. In a word-aligned sentence-pair, it is indicated which target words correspond to each source word. An example of this (which takes French as the source language) is given in the second extract.

✍**11.** A **Sentence-Aligned Corpus**:

Often, in the textile industry, businesses close their plant in Montreal to move to the Eastern Townships.
Dans le domaine du textile souvent, dans Montréal, on ferme et on va s' installer dans les Cantons de l' Est.

There is no legislation to prevent them from doing so, for it is a matter of internal economy.
Il n' y a aucune loi pour empêcher cela, c' est de la régie interne.

But then, in the case of the Gulf refinery it is different : first of all, the Federal Government asked Petro-Canada to buy everything, except in Quebec.
Mais là, la différence entre la Gulf... c' est différent parce que la vente de la raffinerie Gulf: premièrement, le gouvernement fédéral a demandé à Petro-Canada de tout acheter, sauf le Québec.

That is serious.
C'est grave.

✍**12. Word Aligned Corpus**:

The Federal Government asked Petro-Canada to buy everything.
Le(1) gouvernement(3) fédéral(2) a demandé(4) à Petro-Canada(5) de(6) tout(8) acheter(7).

✍**13.** The numbers after the source words indicate the string position of the corresponding target word or words. If there is no target correspondence, then no bracketted numbers appear after the source word (e.g. a in a *demand´e*). If more than one word in the target corresponds, then this is also indicated. The fertility of a source word is the number of words corresponding to it in the target string. For example, the **fertility** of asked with English as source language is 2, since it aligns with a *demand´e*. A third notion is that of **distortion** which refers to the fact that source words and their target correspondences do not necessarily appear in the same string position (compare *tout acheter* and buy everything, for example).

✍**14.** The parameters which must be calculated from the bilingual sentence aligned corpus are:

- the **fertility probabilities** for each source word (i.e. the likelihood of it translating as one, two, three, etc, words respectively),
- the **word-pair** or **translation possibilities** for each word in each language
- the set of **distortion probabilities** for each source and target position.

Short Answer Items & Tests

1.2 Short Answer Items

1. The task of an automatic ……. is to take a formal grammar and a sentence and apply the grammar to the sentence in order to check that it is indeed grammatical.

2. The reasons for ……. languages' superior MT performance are easy to understand. First, the restricted vocabulary means that fewer words need to be added to the MT system dictionaries and more effort can be put into getting the entries which are required right. Second, the grammar component of the system can be tailored to handle all and only those constructions which are licensed by the ……. language specification, a specification which excludes the most difficult and ambiguous constructions anyway.

3. In MT, it is quite common to use the term ……. rather loosely to refer not just to a specialized language, but to its use in a particular type of text, or with a particular ……. purpose.

4. The term …….-based MT has come to describe a …….-based system displaying extensive semantic and pragmatic knowledge of a domain, including an ability to reason, to some limited extent, about concepts in the domain.

5. With respect to MT, the term '……. approaches' can be understood in a narrow sense to refer to approaches which try to do away with explicitly formulating linguistic knowledge, or in a broad sense to denote the application of probablistically based techniques to parts of the MT task.

1.3 Answers

1) parser	**2)** controlled, controlled
3) sublanguage, communicative	**4)** knowledge, rule
5) statistical	

☙☆❧ 1.4 Tests ☙☆❧

✍ Select the best choice.

1. The philosopher Bar-Hillel in a *1959* report argued that fully automatic, high quality, MT (FAHQMT) was ……. .

a) unattainable just at present, but not in principle
b) feasible, not in the past but at present
c) practical in principle but not just at present
d) impossible, not just at present, but in principle

2. Knowledge can be manipulated automatically in two stages: First, we will look at what is called ……., or ……. . This is the process of taking an input string of expressions, and producing representations of the kind we have seen in the previous section. Second, we will look at ……., or ……., which is the reverse process—taking a representation, and producing the corresponding sentence.

a) analysis, generation, synthesis, parsing
b) analysis, synthesis, generation, parsing
c) analysis, parsing, synthesis, generation
d) non of the above is correct

3. A ……. engine generally preserves the surface order of the ……. language and directly re-uses it—with modifications where appropriate—to order the ……. language words.

a) Transformer, source, target
b) LK, source, target
c) Transformer, target, source
d) LK, target, source

4. An ……. engine extracts all the information it can from the source word order and recodes this information in a more or less abstract representation.

a) LK
b) Transformer
c) either a or b
d) neither a nor b

5. From the engine manufacturer's point of view, the ……. architecture has the advantage that it accepts anything that is given to it (though the translations it produces are another matter). The ……. architecture is at a disadvantage here: because it thinks it knows something about the languages involved, it tends to think that anything it doesn't know isn't language and hence unacceptable. As a consequence, a pure ……. engine during its development phase tends to grind to a halt on anything unusual, or even on something quite common which the developer has forgotten to include.

a) LK, LK, transformer
b) transformer, transformer, LK
c) LK, transformer, transformer
d) transformer, LK, LK

6. ……. **refer to cases where one language has to use a phrase to express what another language expresses in a single word.**
a) grammatical gaps
b) syntactic transformations
c) Lexical holes
d) semantic modulations

7. It is useful to think of the kind of knowledge that systems are equipped with as being of three kinds: linguistic knowledge which is independent of context, ……. **knowledge; linguistic knowledge which relates to the context (e.g. of earlier utterances), sometimes called** ……. **knowledge; common sense, general, non-linguistic knowledge about the real world, which we will call** ……. **knowledge.**
a) semiotic, syntactic, pragmatic
b) semantic, pragmatic, real world
c) semantic, syntactic, paradigmatic
d) syntactic, pragmatic, syntagmatic

8. The term ……. **refers to the** ……. **language used in certain fields of knowledge, for example, the language of weather reports, stockmarket reports, the language of some kinds of medical discussion, the language of aeronautical engineering.**
a) sublanguage, specialized
b) sublanguage, generalized
c) meta-language, specialized
d) meta-language, generalized

9. By measuring intelligibility we get only a partial view of translation quality. A highly intelligible output sentence need not be a correct translation of the source sentence. It is important to check whether the meaning of the source language sentence is preserved in the translation. This property is called ……. **.**
a) Clarity or Brevity
b) Precision or Conformity
c) Accuracy or Fidelity
d) Conformity or Clarity

10. A pure example-based approach would use …… .
a) only example phrases
b) no grammar rules at all
c) neither of the above is correct
d) both a and b are correct

☙✷❧ 1.5 Answer key ☙✷❧

	a	b	c	d		a	b	c	d
1				×	2			×	
3	×				4	×			
5				×	6			×	
7		×			8	×			
9			×		10				×

Book ❷

Speech and Language Processing

Daniel Jurafsky & James H. Martin

☙❁❧ 2.1 Notes ☙❁❧

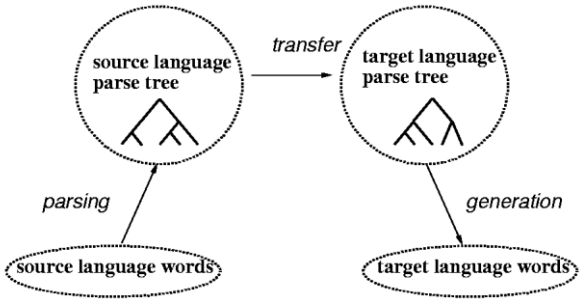
💣 Only the main points in *Chapter 25* of the Book are summarized.

Chapter 25
MACHINE TRANSLATION

✍ **1.** Even when languages differ, these differences often have systematic structure. The study of systematic cross-linguistic similarities and differences is called typology (Croft (1990),

✍ **2.** On **transfer model**, MT involves three phases: **analysis**, **transfer**, and **generation**, where transfer bridges the gap between the output of the source language parser and the input to the target language generator.

✍ **3.** The **transfer architecture** for Machine Translation:



✍ **4.** A simple **transformation** that **reorders** adjectives and nouns:

✍ **5.** In general, **syntactic transformations** are operations that map from one tree structure to another.

6. The process of finding target language equivalents for the content words of the input is called **lexical transfer**. The foundation of **lexical transfer** is dictionary lookup in a cross-language dictionary.

7. One **problem** with the **transfer model** is that it requires a distinct set of transfer rules for each pair of languages.

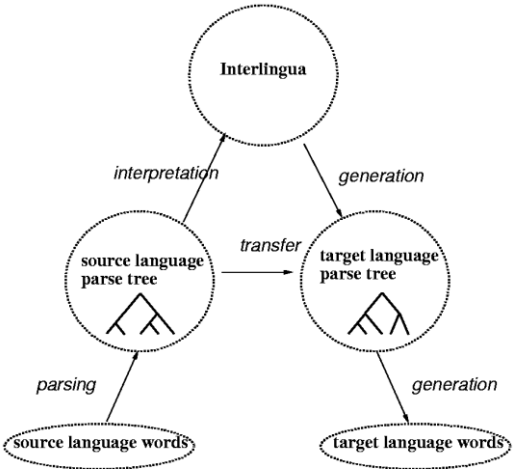
8. The **transfer model** treats translation as a process of altering the structure and words of an input sentence to arrive at a valid sentence of the target language. An alternative to is to treat translation as a process of extracting the meaning of the input and then expressing that meaning in the target language. If this can be done, a MT system can do without *contrastive knowledge*, merely relying on the same syntactic and semantic rules used by a standard interpreter and generator for the language. The amount of *knowledge* needed is then proportional to the number of languages the system handles, rather than to the square, or so the argument goes. This scheme presupposes the existence of a meaning representation, or **interlingua**, in a language-independent canonical form,

9. Interlingual representation of *there was an old man gardening*:

$$\begin{bmatrix} \text{EVENT} & \text{GARDENING} \\ \text{AGENT} & \begin{bmatrix} \text{MAN} & \\ \text{NUMBER} & \text{SG} \\ \text{DEFINITENESS} & \text{INDEF} \end{bmatrix} \\ \text{ASPECT} & \text{PROGRESSIVE} \\ \text{TENSE} & \text{PAST} \end{bmatrix}$$

10. The **interlingua** idea has implications not only for **syntactic transfer** but also for **lexical transfer**. The idea is to avoid explicit descriptions of the relations between source language words and target language words, in favor of mapping via **concepts**, that is, language-independent elements of the ontology.

11. Diagram Suggesting the Relation Between the **Transfer** and **Interlingua** Models, generally credited to Vauqois:



12. Six Stages for a **Direct MT** System for Japanese to English:

Stage	Action
1.	morphological analysis
2.	lexical transfer of content words
3.	various work relating to prepositions
4.	SVO rearrangements
5.	miscellany
6.	morphological generation

13. An Example of Processing in a **Direct System**:

Input:	watashihatsukuenouenopenwojonniageta.
After stage 1:	watashi ha tsukue no ue no pen wo jon ni ageru PAST.
After stage 2:	I ha desk no ue no pen wo John ni give PAST.
After stage 3:	I ha pen on desk wo John to give PAST.
After stage 4:	I give PAST pen on desk John to.
After stage 5:	I give PAST the pen on the desk to John.
After stage 6:	I gave the pen on the desk to John.

14. In the **direct model**, all the processing involving analysis of one specific problem (prepositions for example) is handled in one stage, including analysis, transfer, and generation aspects. The advantage of this is that solving specific problems one at a time may be more tractable. On the other hand, it can be advantageous to organize processing into larger modules (analysis, transfer, synthesis) if there is synergy among all the various individual analysis problems, or among all the individual generation problems, etc.

15. A second characteristic of **direct systems** is that lexical transfer may be more procedural. Lexical transfer procedures may eclectically look at the syntactic classes and semantic properties of neighboring words and

dependents and heads, as seen in the decision-tree-like procedure for translating much and many into Russian in Figure 21.10.

✍**16.** A third characteristic of **direct models** is that they tend to be conservative, to only reorder words when required by obvious ungrammaticality in the result of direct word-for-word substitution. In particular, direct systems generally do lexical transfer before syntactic processing.

✍**17.** Perhaps the key characteristic of **direct models** is that they do without complex structures and representations. In general, they treat the input as a string of words (or morphemes), and perform various operations directly on it—replacing source language words with target language words, re-ordering words, etc.—to end up with a string of symbols in the target language.

✍**18.** In practice, of course, working MT systems tend to be combinations of the **direct**, **transfer**, and **interlingua** methods. But of course syntactic processing is not an all-or-nothing thing. Even if the system does not do a full parse, it can adorn its input with various useful syntactic information, such as part of speech tags, segmentation into clauses or phrases, dependency links, and bracketings. Many systems that are often characterized as direct translation systems also adopt various techniques generally associated with the transfer and interlingua approaches (Hutchins and Somers, 1992).

✍**19.** We can model the **goal** of translation as the production of an output that maximizes some value function that represents the importance of both **faithfulness** and **fluency**. If we chose the product of fluency and faithfulness as our quality metric, we can formalize the translation problem as: *(T is the target-language-sentence and S the source-language-sentence)*

$$\text{best-translation } \hat{T} = \text{argmax}_T \text{ fluency(T) faithfulness(T,S)}$$

$$\text{best-translation } \hat{T} = \text{argmax}_T \text{ P(T) } P(S|T)$$

✍**20.** We need to do three things: quantify fluency, P(T), quantify faithfulness, P(S|T) and create an algorithm to find the sentence that maximizes the product of these two things. There is an innovation here. In the **transfer**, **interlingua**, and **direct models**, each step of the process made some adjustment to the input sentence to make it closer to a *fluent* TL sentence, while obeying the constraint of not changing the meaning too much. In those models the process is fixed, in that there is no flexibility to trade-off a modicum of *faithfulness* for a smidgeon of naturalness, or conversely, based on the specific input sentence at hand. This new model, sometimes called the **statistical model of translation** allows exactly that.

21. It has also become apparent that MT systems do better if the dictionaries include not only **words** but also **idioms**, **fixed phrases**, and even **frequent clauses** and **sentences**. Such data can sometimes be extracted automatically from corpora. Moreover, in some situations it may be valuable to do this on-line, at translation time, rather than saving the results in a dictionary—this is they key idea behind **Example-based Machine Translation** (Sumira and Iida, 1991; Brown, 1996).

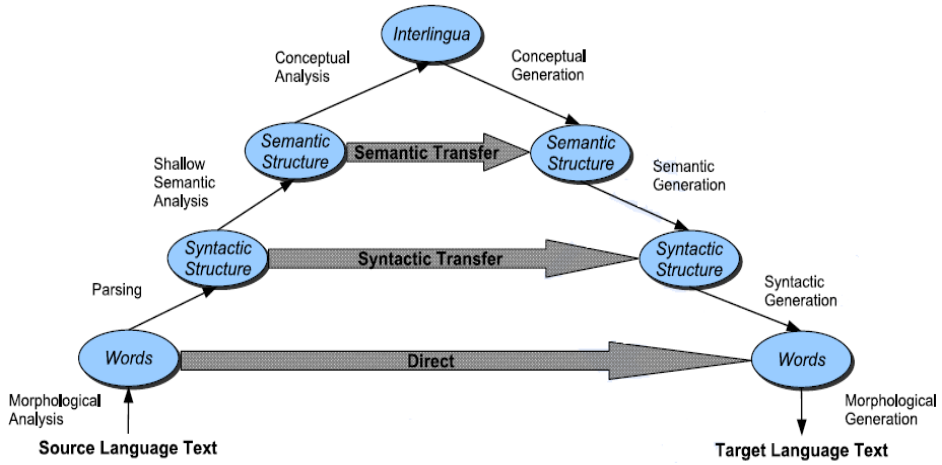
22. In **direct translation**, we proceed word-by-word through the source language text, translating each word as we go.

23. Direct translation uses a large bilingual dictionary, each of whose entries is a small program with the job of translating one word.

24. In **transfer approaches**, we first parse the input text, and then apply rules to transform the source language parse structure into a target language parse structure. We then generate the target language sentence from the parse structure.

25. In **interlingua approaches**, we analyze the source language text into some abstract meaning representation, called an interlingua. We then generate into the target language from this interlingual representation.

26. The **Vauquois triangle** shows the increasing depth of analysis required (on both the analysis and generation end) as we move from the **direct** approach through **transfer** approaches, to **interlingual** approaches. In addition, it shows the decreasing amount of transfer knowledge needed as we move up the triangle, from huge amounts of transfer at the **direct** level (almost all knowledge is transfer knowledge for each word) through **transfer** (transfer rules only for parse trees or thematic roles) through **interlingua** (no specific transfer knowledge).

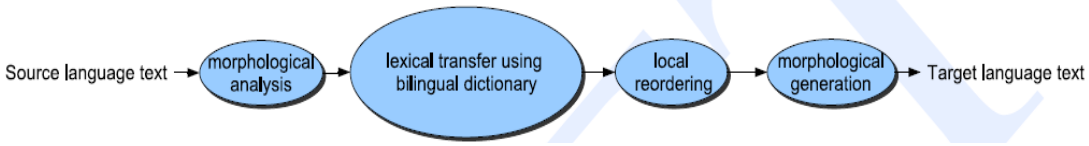


27. In **direct translation**, we proceed word-by-word through the source language text, translating each word as we go. We make use of no intermediate structures, except for shallow morphological analysis; each source word is directly mapped onto some target word.

28. Direct translation is thus based on a large bilingual dictionary; each entry in the dictionary can be viewed as a small program whose job is to translate one word. After the words are translated, simple reordering rules can apply, for example for moving adjectives after nouns when translating from English to French.

29. The guiding intuition of the **direct approach** is that we translate by incrementally transforming the source language text into a target language text. While the pure **direct approach** is no longer used, this transformational intuition underlies all modern systems, both statistical and non-statistical.

30. ***Direct machine translation***: The major component, indicated by size here, is the bilingual dictionary:



31. The four steps outlined in the above figure would proceed as shown in the following figure. **Step 2** presumes that the bilingual dictionary has the phrase *dar una bofetada* a as the Spanish translation of English slap. The local reordering **step 3** would need to switch the adjective-noun ordering from green witch to *bruja verde*. And some combination of ordering rules and the dictionary would deal with the negation and past tense in English didn't.

Input:	Mary didn't slap the green witch
After 1: Morphology	Mary DO-PAST not slap the green witch
After 2: Lexical Transfer	Maria PAST no dar una bofetada a la verde bruja
After 3: Local reordering	Maria no dar PAST una bofetada a la bruja verde
After 4: Morphology	Maria no dió una bofetada a la bruja verde

32. While the **direct approach** can deal with our simple Spanish example, and can handle single-word **reorderings**, it has no parsing component or indeed any knowledge about phrasing or grammatical structure in the source or target language. It thus cannot reliably handle longer-distance **reorderings**, or those involving phrases or larger structures. Even more complex **reorderings** occur when we translate from ***SVO*** to ***SOV*** languages, as we see in the English-Japanese example from Yamada and Knight (2002):

He adores listening to music

kare	*ha*	*ongaku*	*wo*	*kiku*	*no*	*ga*	*daisuki*	*desu*
he		music	to	listening			adores	

33. These examples suggest that the **direct approach** is too *focused on individual words*, and that in order to deal with real examples we'll need to add **phrasal** and **structural knowledge** into our MT models.

34. Languages differ systematically in **structural** ways. One strategy for doing MT is to translate by a process of overcoming these differences, altering the **structure** of the input to make it conform to the **rules** of the target language. This can be done by applying **contrastive knowledge**, that is, knowledge about the differences between the two languages. Systems that use this strategy are said to be based on the **transfer model**.

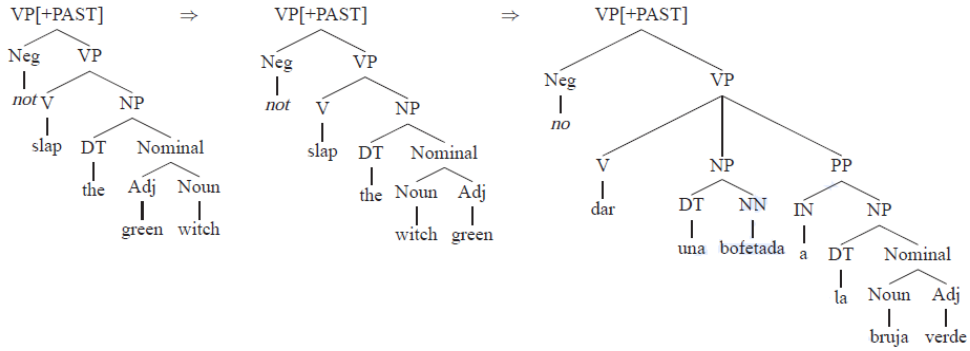
35. The **transfer model** presupposes a parse of the source language, and is followed by a generation phase to actually create the output sentence. Thus, on this model, MT involves three phases: analysis, transfer, and generation, where transfer bridges the gap between the output of the source language parser and the input to the target language generator.

36. It is worth noting that a **parse** for MT may *differ* from parses required for other purposes. For example, suppose we need to translate John saw the girl with the binoculars into French. The parser does not need to bother to figure out where the prepositional phrase attaches, because both possibilities lead to the same French sentence.

37. Once we have **parsed the source** language, we'll need rules for **syntactic transfer** and **lexical transfer**. The syntactic transfer rules will tell us how to modify the source parse tree to resemble the target parse tree.

38. A simple transformation that reorders ***adjectives*** and ***nouns***:

39. The **transfer approach** and this **rule** can be applied to our example *Mary did not slap the green witch*. Besides this **transformation rule**, we'll need to assume that the morphological processing figures out that didn't is composed of do-PAST plus not, and that the parser attaches the PAST feature onto the VP. Lexical transfer, via lookup in the bilingual dictionary, will then remove do, change not to no, and turn slap into the phrase *dar una bofetada a*, with a slight rearrangement of the parse tree.



40. In addition to syntactic transformations, **transfer-based systems** need to have lexical transfer rules. Lexical transfer is generally based on a bilingual dictionary, just as for direct MT. The dictionary itself can also be used to deal with problems of lexical ambiguity. For example the English word home has many possible translations in German, including *nach Hause* (in the sense of going home) *Heim* (in the sense of a home game), *Heimat* (in the sense of homeland, home country, or spiritual home), and *zu Hause* (in the sense of being at home). In this case, the phrase at home is very likely to be translated *zu Hause*, and so the bilingual dictionary can list this translation idiomatically.

41. Commercial MT systems tend to be combinations of the **direct** and **transfer approaches**, using rich *bilingual dictionaries*, but also using taggers and parsers. The Systran system, for example, as described in Hutchins and Somers (1992), Senellart et al. (2001), has three components.

First is a **shallow analysis** stage, including:

- ✺ morphological analysis and part of speech tagging
- ✷ chunking of NPs, PPs, and larger phrases
- ◎ shallow dependency parsing (subjects, passives, head-modifiers)

Next is a **transfer phase**, including:

- ✪ translation of idioms,
- ❄ word sense disambiguation
- ❀ assigning prepositions based on governing verbs

Finally, in the **synthesis stage**, the system:

- ⌘ applies a rich bilingual dictionary to do lexical translation
- ▣ deals with reorderings
- ✠ performs morphological generation

42. Like the **direct** system, the **Systran** system relies for much of its processing on the **bilingual dictionary**, which has lexical, syntactic, and semantic knowledge.

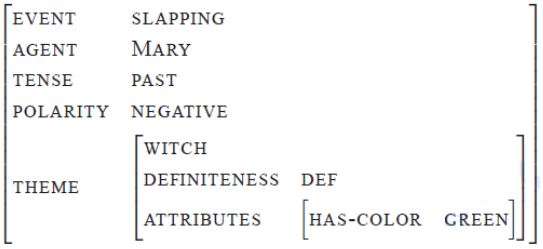
43. Like a **direct** system, **Systran** does **reordering** in a post-processing step. Like a transfer system, many of the steps are informed by *syntactic* and *shallow semantic* processing of the source language.

44. One problem with the **transfer model** is that it requires a distinct set of transfer rules for each pair of languages. This is clearly suboptimal for translation systems employed in many-to-many multilingual environments like the European Union. This suggests a different perspective on the nature of translation. Instead of directly transforming the words of the source language sentence into the target language, the **interlingua** intuition is to treat translation as a process of extracting the meaning of the input and then expressing that meaning in the target language.

45. If this could be done, an MT system could do **without contrastive knowledge**, merely relying on the same syntactic and semantic rules used by a standard interpreter and generator for the language. The amount of knowledge needed would then be proportional to the number of languages the system handles, rather than to the square.

46. This scheme presupposes the existence of a **meaning representation**, or **interlingua**, in a language-independent canonical form. The idea is for the **interlingua** to represent all sentences that mean the "**same**" thing in the same way, regardless of the language they happen to be in. Translation in this model proceeds by performing a **deep semantic analysis** on the input from language X into the **interlingual representation** and generating from the interlingua to language Y.

47. The following figure shows a possible **interlingual representation** for *Mary did not slap the green witch* as a unification-style feature structure. We can create these **interlingual representation** from the source language text using the **semantic analyzer** techniques; using a **semantic role** labeler to discover the AGENT relation between *Mary* and the *slap* event, or the THEME relation between the witch and the *slap* event. We would also need to do **disambiguation** of the noun-modifier relation to recognize that the relationship between *green* and *witch* is the has-color relation, and we'll need to discover that this event has negative polarity (from the word didn't). The **interlingua** thus requires more analysis work than the **transfer model**, which only required **syntactic parsing** (or at most shallow thematic role labeling). But generation can now proceed directly from the **interlingua** with no need for **syntactic transformations**.



48. The **interlingual model** has its own **problems**. For example, in order to translate from Japanese to Chinese the universal interlingua must include concepts such as *ELDER-BROTHER* and *YOUNGER-BROTHER*. Using these same concepts translating from German-to-English would then require large amounts of unnecessary disambiguation. Furthermore, doing the extra work involved by the **interlingua** commitment requires exhaustive analysis of the semantics of the domain and formalization into an ontology. Generally this is only possible in relatively simple domains based on a database model, as in the air travel, hotel reservation, or restaurant recommendation domains, where the database definition determines the possible entities and relations. For these reasons, **interlingual systems** are generally only used in sublanguage domains.

49. We can model the goal of translation as the production of an output that maximizes some value function that represents the importance of both ***faithfulness*** and ***fluency***. **Statistical MT** is the name for a class of approaches that do just this, by building probabilistic models of faithfulness and fluency, and then combining these models to choose the most probable translation. If we chose the product of ***faithfulness*** and ***fluency*** as our quality metric, we could model the translation from a source language sentence S to a target language sentence T$\hat{}$ as:

best-translation T$\hat{}$ = argmaxT faithfulness(T,S) fluency(T)

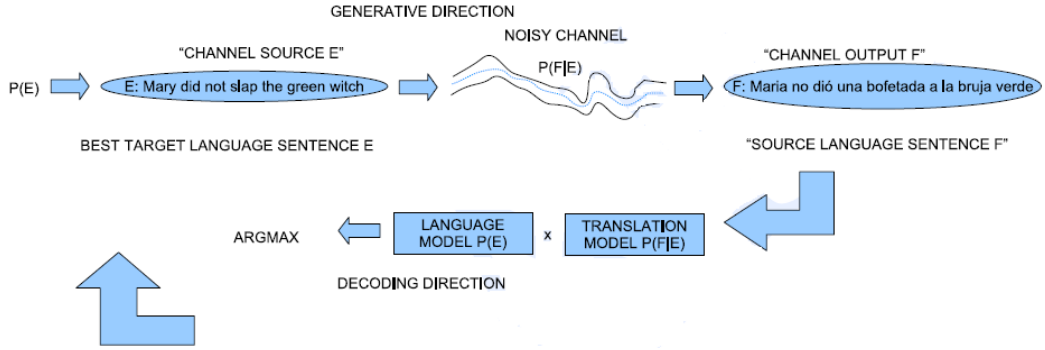
50. This intuitive equation clearly resembles the **Bayesian noisy** channel model for speech. Let's make the analogy perfect and formalize the ***noisy channel model*** for **statistical machine translation**. First of all, for the rest of this chapter, we'll assume we are translating from a foreign language sentence $F = f_1, f_2, ..., f_m$ to English. For some examples we'll use French as the foreign language, and for others Spanish. But in each case we are translating into English (although of course the statistical model also works for translating out of English). In a probabilistic model, the best English sentence $\hat{E} = e_1,e_2, ...,e_l$ is the one whose probability P(E|F) is the highest. As is usual in the ***noisy channel model***, we can rewrite this via **Bayes rule**:

$$\begin{aligned}\hat{E} &= argmax_E P(E|F) \\ &= argmax_E \frac{P(F|E)P(E)}{P(F)} \\ &= argmax_E P(F|E)P(E)\end{aligned}$$

51. We can ignore the denominator P(F) inside the argmax since we are choosing the best English sentence for a fixed foreign sentence F, and hence P(F) is a constant. The resulting noisy channel equation shows that we need two components: **a translation model** P(F|E), and a **language model** P(E).

$$\hat{E} = \underset{E \in \text{English}}{\text{argmax}} \overbrace{P(F|E)}^{\text{translation model}} \overbrace{P(E)}^{\text{language model}}$$

52. Notice that applying the **noisy channel model** to machine translation requires that we think of things backwards, as shown in the following figure. We pretend that the foreign (source language) input F we must translate is a corrupted version of some English (target language) sentence E, and that our task is to discover the hidden (target language) sentence E that generated our observation sentence F.

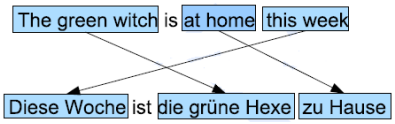


53. The *noisy channel model* of **statistical MT** thus requires three components to translate from a French sentence F to an English sentence E:

- A language model to compute P(E)
- A translation model to compute P(F|E)
- A decoder, which is given F and produces the most probable E

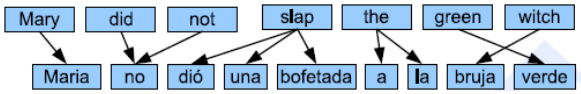
54. Statistical MT systems are based on the same **N-gram language models** as speech recognition and other applications. The **language model** component is monolingual, and so acquiring training data is relatively easy.

55. The job of the **translation model**, given an English sentence E and a foreign sentence F, is to assign a probability that E generates F. While we can estimate these probabilities by thinking about how each individual word is translated, modern statistical MT is based on the intuition that a better way to compute these probabilities is by considering the behavior of phrases. As we see in the following figure, entire phrases often need to be translated and moved as a unit. The intuition of **phrase-based statistical MT** is to use phrases (sequences of words) as well as single words as the fundamental units of translation.

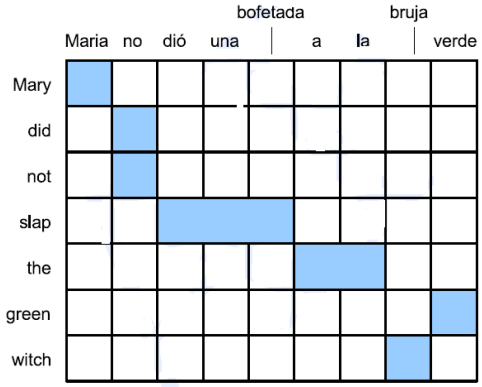


56. All **statistical translation models** are based on the idea of a **word alignment**. A word alignment is a **mapping** between the source words and the target words in a set of parallel sentences.

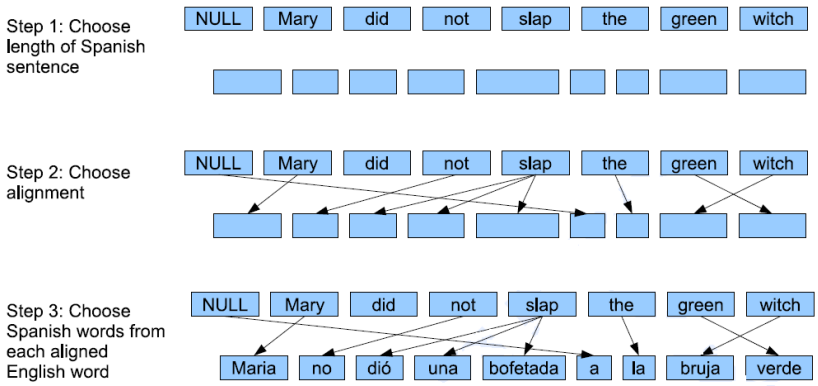
57. A **graphical model representation** of a **word alignment** between the English and Spanish sentences:



58. An **alignment matrix representation** of a word alignment between the English and Spanish sentences:

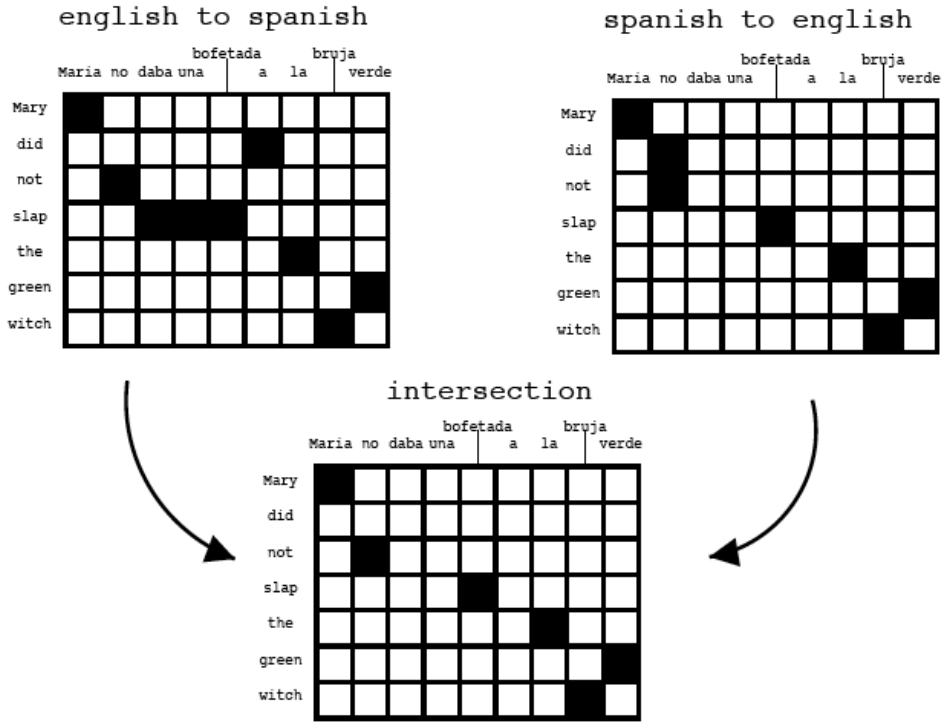


59. The three steps of **IBM Model 1** generating a Spanish sentence and alignment from an English sentence:



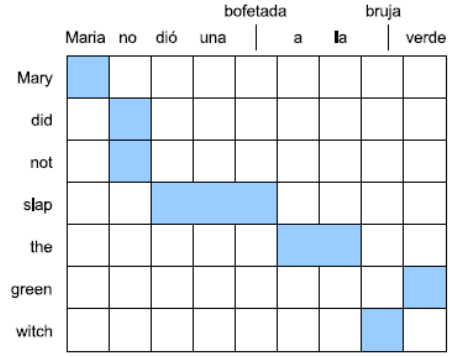
60. All **statistical translation models** are trained using a large **parallel corpus**. A parallel corpus, **parallel text**, or **bitext** is a text that is available in two languages.

61. ***PLACEHOLDER FIGURE:*** (Intersection of English-to-Spanish and Spanish-to-English alignments to produce a high-precision alignment. Alignment can then be expanded with points from both alignments to produce an alignment like that shown in the figure 2)



62. ***Figure 2.*** A better **phrasal alignment** for *the green witch* sentence, computed by starting with the intersection alignment in the previous

figure and adding points from the union alignment, using the algorithm of Och and Ney (2003).

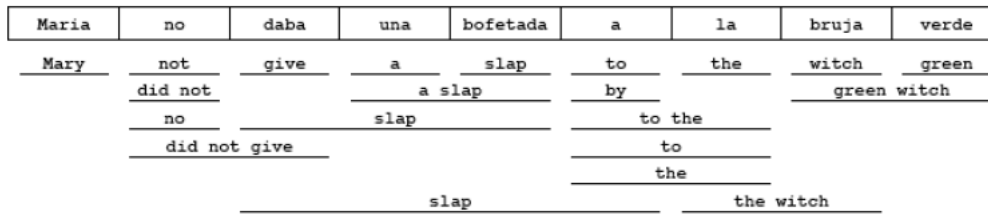


63. The job of the **decoder** is to take a foreign (Spanish) source sentence F and produce the best (English) translation E according to the product of the translation and language models:

$$\hat{E} = \underset{E \in \text{English}}{\text{argmax}} \overbrace{P(F|E)}^{\text{translation model}} \overbrace{P(E)}^{\text{language model}}$$

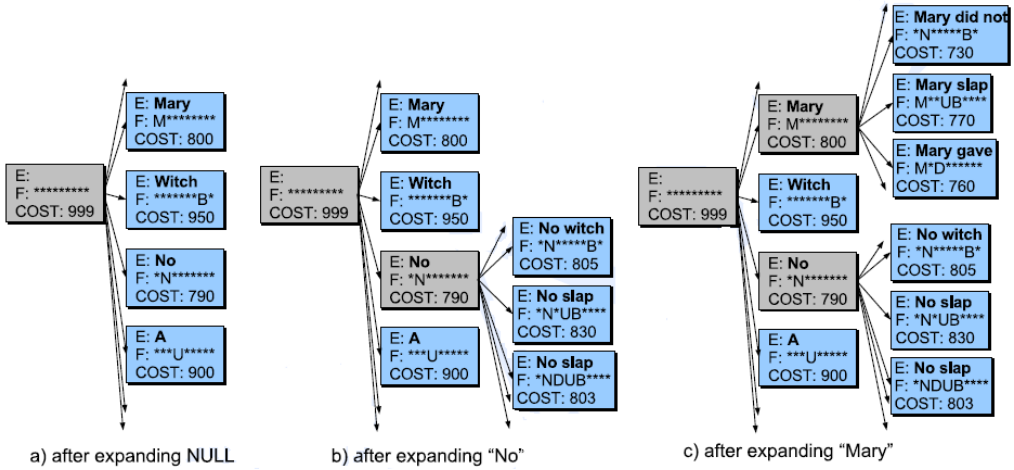
64. Finding the sentence which maximizes the translation and language model probabilities is a search problem, and **decoding** is thus a kind of search. **Decoders** in MT are based on best-first search, a kind of heuristic or informed search; these are search algorithms that are informed by knowledge from the problem domain. Best-first search algorithms select a node n in the search space to explore based on an evaluation function f (n). MT decoders are variants of a specific kind of best-first search called A* search. A* search was first implemented for machine translation by IBM (Brown et al., 1995), based on IBM's earlier work on A* search for speech recognition (Jelinek, 1969). *[stack or A* decoding]*

65. ***PLACEHOLDER FIGURE***: The lattice of possible English translations for words and phrases in a particular sentence F, taken from the entire aligned training set:



66. Three stages in stack decoding of *Maria no di´o una bofetada a la bruja verde* (simplified by assuming a single stack and no pruning). The

nodes in blue, on the fringe of the search space, are all on the stack, and are **open** nodes still involved in the search. Nodes in gray are **closed** nodes which have been popped off the stack.



67. Now let's walk informally through the stack decoding example in the above Figure, producing an English translation of *Mary di´o una bofetada a la bruja verde* left to right. For the moment we'll make the simplifying assumption that there is a single stack, and that there is no pruning. We start with the null hypothesis as the initial **search state**, in which we have selected no Spanish words and produced no English translation words. We now **expand** this hypothesis by choosing each possible source word or phrase which could generate an English sentence-initial phrase.

68. The above ***Figure (a)*** shows this first ply of the search. For example the top state represents the hypothesis that the English sentence starts with Mary, and the Spanish word *Maria* has been covered (the asterisk for the first word is marked with an M). Each state is also associated with a cost, discussed below. Another state at this ply represents the hypothesis that the English translation starts with the word *No*, and that Spanish no has been covered. This turns out to be the lowest-cost node on the queue, so we pop it off the queue and push all its expansions back on the queue. Now the state *Mary* is the lowest cost, so we expand it; *Mary did not* is now the lowest cost translation so far, so will be the next to be expanded. We can then continue to expand the search space until we have states (hypotheses) that cover the entire Spanish sentence, and we can just read off an English translation from this state. We mentioned that each state is associated with a cost which, as we'll see below, is used to guide the search, The cost combines the current cost with an estimate of the **future cost**. The **current cost** is the total probability of the phrases

that have been translated so far in the hypothesis, i.e. the product of the translation, distortion, and language model probabilities.

69. Broadly speaking, we attempt to evaluate translations along two dimensions, corresponding to the **fidelity** and **fluency**.

70. The most **accurate** evaluations use **human raters** to evaluate each translation along each dimension. For example, along the dimension of *fluency*, we can ask how *intelligible*, how *clear*, how *readable*, or how *natural* is the MT output (the target translated text).

71. There are two broad ways: One method is to give the raters a **scale**, for example from 1 (totally unintelligible) to 5 (totally intelligible), and ask them to rate each sentence or paragraph of the MT output. We can use distinct scales for any of the aspects of fluency, such as clarity, naturalness, or style.

72. The second class of methods relies **less** on the **conscious decisions** of the participants. For example, we can measure the time it takes for the raters to read each output sentence or paragraph. Clearer or more fluent sentences should be faster or easier to read. We can also measure fluency with the cloze task (Taylor, 1953, 1957).

73. The **cloze task** is a metric used often in psychological studies of reading. The rater sees an output sentence with a word replaced by a space (for example, every 8th word might be deleted). Raters have to guess the identity of the missing word. Accuracy at the cloze task, i.e. average success of raters at guessing the missing words, generally correlates with how intelligible or natural the MT output is.

74. Two common aspects of **fidelity** which are measured are **adequacy** and **informativeness**.

75. The **adequacy** of a translation is whether it contains the information that existed in the original. We measure adequacy by using raters to assign scores on a scale.

- If we have bilingual raters, we can give them the source sentence and a proposed target sentence, and rate, perhaps on a 5-point scale, how much of the information in the source was preserved in the target.
- If we only have monolingual raters, but we have a good human translation of the source text, we can give the monolingual raters the human reference translation and a target machine translation, and again rate how much information is preserved.

76. The **informativeness** of a translation is a task-based evaluation of whether there is sufficient information in the MT output to perform some task. For example we can give raters multiple-choice questions about the content of the material in the source sentence or text. The raters answer

these questions based only on the MT output. The percentage of correct answers is an informativeness score.

✍**77.** Another set of metrics attempt to judge the overall **quality** of a translation, **combining fluency** and **fidelity**. For example, the typical evaluation metric for MT output to be post-edited is the edit cost of post-editing the MT output into a good translation. For example, we can measure the number of words, the amount of time, or the number of keystrokes required for a human to correct the output to an acceptable level.

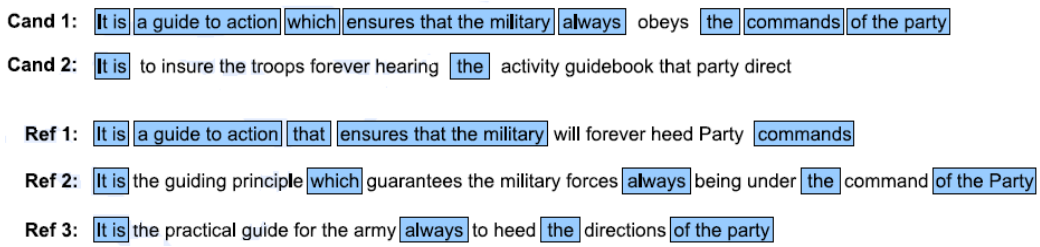
✍**78.** While **humans** produce the best evaluations of machine translation output, running a human evaluation can be very **time-consuming**, taking days or even weeks. It is useful to have an **automatic metric** that can be run relatively frequently to quickly evaluate potential system improvements. In order to have such convenience, we would be willing for the metric to be much worse than human evaluation, as long as there was some correlation with human judgments.

✍**79.** In fact there are a number of such heuristic methods, such as **Bleu**, **NIST**, **TER**, **Precision and Recall**, and **METEOR**. The intuition of these automatic metrics derives from Miller and Beebe-Center (1958), who pointed out that **a good MT output is one which is very similar to a human translation**.

✍**80.** For each of these metrics, we assume that we already have one or **more human translations** of the relevant sentences. Now given an MT output sentence, we compute the translation closeness between the MT output and the human sentences. An MT output is ranked as better if on average it is **closer to the human translations**. The metrics differ on what counts as 'translation closeness'.

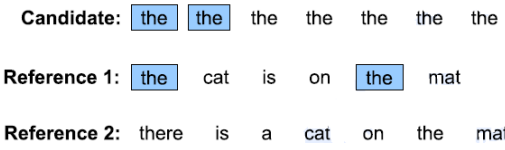
✍**81.** In the field of automatic speech recognition, the metric for 'transcription closeness' is word error rate, which is the minimum edit distance to a human transcript. But in translation, we can't use the same word error rate metric, because there are many possible translations of a source sentence; a very good MT output might look like one human translation, but very unlike another one. For this reason, most of the metrics judge an MT output by comparing it to **multiple human translations**.

✍**82.** The following figure shows an intuition, from **two candidate** translations of a Chinese source sentence, shown with **three reference** human translations of the source sentence. Note that Candidate 1 shares many more words (shown in blue) with the reference translations than Candidate 2. *{Intuition for Bleu: one of two candidate translations of a Chinese source sentence shares more words with the reference human translations.}*



83. A **basic unigram precision metric** would be to count the number of words in the candidate translation (MT output) that occur in some reference translation, and divide by the total number of words in the candidate translation. If a candidate translation had 10 words, and 6 of them occurred in at least one of the reference translations, we would have a precision of 6/10 = 0.6.

84. Alas, there is a ***flaw*** in using simple precision: *it rewards candidates that have extra repeated words.* The following figure shows an example of a pathological candidate sentence composed of multiple instances of the single word the. Since each of the 7 (identical) words in the candidate occurs in one of the reference translations, the unigram precision would be 7/7!



85. In order to avoid this problem, Bleu uses a modified **N-gram precision metric**. We first count the **maximum** number of times a word is used in any single reference translation. The count of each candidate word is then clipped by this maximum reference count. Thus the modified unigram precision in the example in Fig. 25.32 would be 2/7, since Reference 1 has a maximum of 2 ***the***s.

86. Going back to Chinese example in the previous figure, Candidate 1 has a modified unigram precision of 17/18, while Candidate 2 has one of 8/14. We compute the **modified precision** similarly for higher order N-grams as well. The **modified bigram** precision for Candidate 1 is 10/17, and for Candidate 2 is 1/13.

87. To compute a score over the whole test set, Bleu first computes the *N-gram matches* for each sentence, and add together the clipped counts over all the candidates sentences, and divide by the total number of candidate N-grams in the test set. The modified precision score is thus:

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n\text{-}gram \in C} \text{Count}_{\text{clip}}(n\text{-}gram)}{\sum_{C' \in \{Candidates\}} \sum_{n\text{-}gram' \in C'} \text{Count}(n\text{-}gram')}$$

✍ **88.** Bleu uses unigram, bigrams, trigrams, and often quadrigrams; it combines these modified N-gram precisions together by taking their geometric mean. In addition, Bleu adds a further penalty to **penalize** candidate translations that are **too short**.

✍ **89.** Consider the candidate translation of the, compared with References 1-3 in the above figure. Because this candidate is so **short**, and all its words appear in some translation, its modified unigram precision is inflated to 2/2. Normally we deal with these problems by combining precision with recall. But we can't use ***recall*** over multiple human translations, since recall would require (incorrectly) that a good translation must contain contains lots of N-grams from every translation. Instead, Bleu includes a **brevity penalty** over the whole corpus. Let c be the total length of the candidate translation corpus. We compute the effective reference length r for that corpus by summing, for each candidate sentence, the lengths of the best matches. The brevity penalty is then an exponential in r/c. In summary:

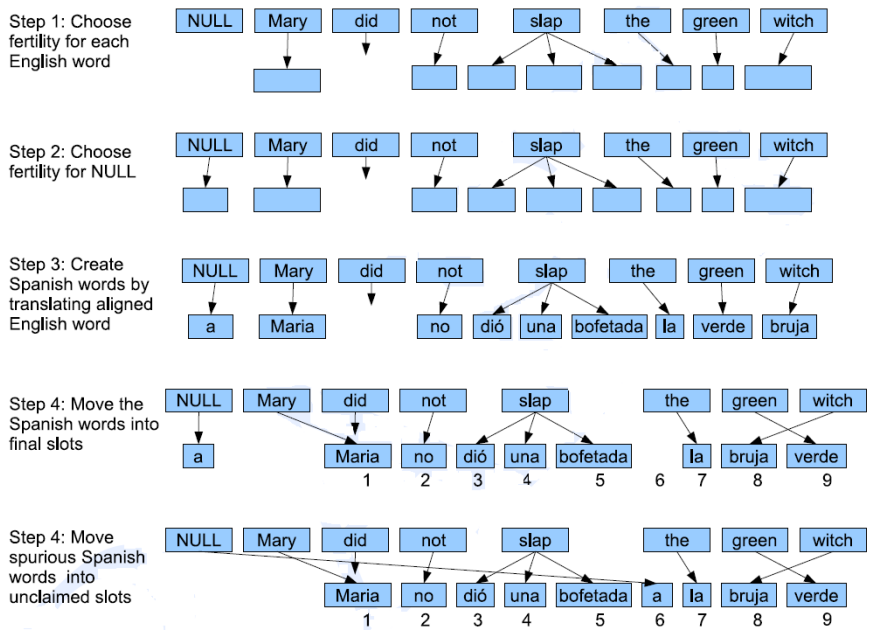
$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

$$\text{Bleu} = \text{BP} \times \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log p_n\right)$$

✍ **90.** While **automatic metrics** like Bleu (or NIST, METEOR, etc) have been very useful in quickly evaluating potential system improvements, and match human judgments in many cases, they have certain limitations that are important to consider. First, many of them focus on very local information. Consider slightly moving a phrase in the previous figure slightly to produce a candidate like: *Ensures that the military it is a guide to action which always obeys the commands of the party*. This sentence would have an identical **Bleu** score to Candidate 1, although a human rater would give it a lower score. Furthermore, the automatic metrics probably do poorly at comparing systems that have radically different architectures. Thus **Bleu**, for example, is known to perform poorly (i.e. not agree with human judgments of translation quality) when evaluating the output of commercial systems like **Systran** against N-gram-based statistical systems, or even when evaluating human-aided translation against machine translation (Callison-Burch et al., 2006).

91. We can conclude that **automatic metrics** are most appropriate when evaluating incremental changes to a single system, or comparing systems with very similar architectures.

92. The five steps of ***IBM Model 3*** generating a Spanish sentence and alignment from an English sentence:



More about MT Evaluation

1. A typical way for **lay people** to assess machine translation quality is to translate from a source language to a target language and **back to the source** language with the same engine. Though intuitively this seems a good method of evaluation, it has been shown that round-trip translation is a, "**poor predictor of quality**". The reason why it is such a poor predictor of quality is reasonably intuitive. A **round-trip translation** is not testing one system, but two systems: the language pair of the engine for translating in to the target language, and the language pair translating back from the target language. Consider the following examples of round-trip translation performed from English to Italian and Portuguese from Somers (2005):

Original text 1	**Select this link to look at our home page.**
Translated	*Selezioni questo collegamento per guardare il nostro Home Page.*

Translated back	Selections this connection in order to watch our Home Page.
Original text 2	**Tit for tat**
Translated	*Melharuco para o tat*
Translated back	Tit for tat

2. In the first example, where the text is translated into Italian then back into English—the English text is significantly garbled, but the Italian is a serviceable translation. In the second example, the text translated back into English is perfect, but the Portuguese translation is meaningless. While **round-trip translation** may be useful to generate a "surplus of fun," the methodology is deficient for serious study of machine translation quality.

3. One of the constituent parts of the **ALPAC report** was a study comparing different levels of human translation with machine translation output, using human subjects as judges. The human judges were specially trained for the purpose. The evaluation study compared an MT system translating from Russian into English with human translators, on two variables. The variables studied were "**intelligibility**" and "**fidelity**".

4. Intelligibility was a measure of how "**understandable**" the sentence was, and was measured on a scale of 1—9. Fidelity was a measure of how much information the translated sentence retained compared to the original, and was measured on a scale of 0—9. Each point on the scale was associated with a textual description. For example, 3 on the intelligibility scale was described as "Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence".

5. Intelligibility was measured without reference to the original, while **fidelity** was measured indirectly. The translated sentence was presented, and after reading it and absorbing the content, the original sentence was presented. The judges were asked to rate the original sentence on informativeness. So, the more ***informative*** the original sentence, the ***lower*** the quality of the translation.

6. The study showed that the variables were highly correlated when the human judgment was averaged per sentence. The variation among raters was small, but the researchers recommended that at the very least, three or four raters should be used. The evaluation methodology managed to separate translations by humans from translations by machines with ease. The study concluded that, "highly reliable assessments can be made of the quality of human and machine translations".

- **7. BLEU** was one of the first metrics to report high correlation with human judgments of quality. The metric is currently one of the most popular in the field. The central idea behind the metric is that "the closer a machine translation is to a professional human translation, the better it is".
- **8. BLEU** calculates scores for individual segments, generally sentences—then **averages** these scores over the whole corpus for a final score. It has been shown to correlate highly with human judgments of quality at the corpus level.
- **9. BLEU** uses a modified form of precision to compare a candidate translation against multiple reference translations. The metric modifies simple precision since machine translation systems have been known to generate more words than appear in a reference text.
- **10.** The **NIST** metric is based on the BLEU metric, but with some alterations. Where BLEU simply calculates n-gram precision adding equal weight to each one, **NIST** also calculates how informative a particular n-gram is. That is to say when a correct n-gram is found, the rarer that n-gram is, the more weight it is given. For example, if the bigram "***on the***" correctly matches, it receives lower weight than the correct matching of bigram "***interesting calculations***," as this is less likely to occur.
- **11. NIST** also differs from BLEU in its calculation of the *brevity penalty*, insofar as small variations in translation length do not impact the overall score as much.
- **12.** The **Word error rate** (WER) is a metric based on the Levenshtein distance, where the Levenshtein distance works at the character level, **WER** works at the word level. It was originally used for measuring the performance of speech recognition systems, but is also used in the evaluation of machine translation.
- **13.** The **Word error rate** is based on the calculation of the number of words that differ between a piece of machine translated text and a reference translation. A related metric is the Position-independent word error rate (**PER**), this allows for re-ordering of words and sequences of words between a translated text and a references translation.
- **14.** The **METEOR** metric is designed to address some of the deficiencies inherent in the BLEU metric. The metric is based on the weighted harmonic mean of unigram precision and unigram recall. The metric was designed after research by Lavie (2004) into the significance of recall in evaluation metrics. Their research showed that metrics based on recall consistently achieved higher correlation than those based on precision alone, cf. BLEU and NIST.

✍ **15. METEOR** also includes some other features not found in other metrics, such as **synonymy** matching, where instead of matching only on the exact word form, the metric also matches on synonyms. For example, the word "good" in the reference rendering as "well" in the translation counts as a match. The metric is also includes a stemmer, which lemmatises words and matches on the lemmatised forms. The implementation of the metric is modular insofar as the algorithms that match words are implemented as modules, and new modules that implement different matching strategies may easily be added.

Short Answer Items & Tests

2.2 Short Answer Items

1. On transfer model, MT involves three phases: analysis, transfer, and generation, where transfer bridges the gap between the output of the source language …… and the input to the target language …… .

2. One problem with the transfer model is that it requires a …… set of transfer …… for each pair of languages.

3. In practice, working MT systems tend to be combinations of the ……, transfer, and …… methods. But of course syntactic processing is not an all-or-nothing thing.

4. The Vauquois triangle shows the increasing depth of analysis required (on both the analysis and generation end) as we move from the …… approach through …… approaches, to …… approaches.

5. Like the …… system, the Systran system relies for much of its processing on the …… dictionary, which has lexical, syntactic, and semantic knowledge.

2.3 Answers

1) parser, generator	**2)** distinct, rules
3) direct, interlingua	**4)** direct, transfer, interlingual
5) direct, bilingual	

☙☆❀ 2.4 Tests ☙☆❀

✍ Select the best choice.

1. On transfer model, MT involves three phases: ……. .

a) synthesis, transmit, creation b) analysis, synthesis, generation
c) analysis, transfer, generation d) synthesis, transfer, generation

2. The interlingua idea has implications ……. **.**

a) not only for syntactic transfer but also for lexical transfer
b) only for syntactic transfer (not for lexical transfer)
c) only for lexical transfer (not for syntactic transfer)
d) neither for syntactic transfer nor for lexical transfer

3. In …… **approaches, we first parse the input text, and then apply rules to transform the source language parse structure into a target language parse structure. We then generate the target language sentence from the parse structure.**

a) intralingua b) transfer
c) direct d) all of the above

4. Perhaps the key characteristic of …… **models is that they do without complex** …… **and** ……**. In general, they treat the input as a string of words (or morphemes), and perform various operations directly on it—replacing source language words with target language words, re-ordering words, etc.—to end up with a string of symbols in the target language.**

a) direct, structures b) direct, representations
c) transfer, analysis, synthesis d) both a and b

5. The …… **idea is to avoid explicit descriptions of the relations between source language words and target language words, in favor of mapping via** ……**, that is, language-**…… **elements of the ontology.**

a) intralingua, structures, dependent
b) interlingua, concepts, independent
c) interlingua, concepts, dependent
d) intralingua, concepts, independent

☙✷❧ 2.5 Answer key ☙✷❧

	a	b	c	d		a	b	c	d
1			×		2	×			
3		×			4				×
5		×							

Book ❸

Machine Translation:

Its Scope and Limits

Yorick Wilks

3.1 Notes

Chapter 1
Introduction

1. **AI** researchers had argued since the mid-seventies that **knowledge-based** systems were the key to MT, as to everything else in intelligent computation, but the problem was that they had failed to deliver **knowledge bases** of sufficient size to test this, thus leaving the **AI** case to rest only on plausible examples, such as "*The soldiers fired at the women and I saw several fall*" (Wilks, 1975a), where one feels that the "*several*" is the *women* not because of any linguistic selection rules or statistical regularities but because of our **knowledge** of how the **world** works.

2. We now have two competing research paradigms in MT, the **symbolic** and the **statistical**, each armed with a set of rock solid examples and arguments but neither able to beat – in open competition – the old commercial legacy system SYSTRAN unaided, systems inherited from the 1970's. Indeed, the MT systems made available by Google are a version of the SYSTRAN system for most languages, but new, statistically-based, systems for Chinese and Arabic.

PART I
MT Past

Chapter 2
Five Generations of MT

✍ **1.** In the following list, the first three correspond very roughly, to what are called the "**generations**" of MT, but to that I think we can add one or two more.

- ✪ **First**, the "brute force" methods for MT, that were thought to have been brought to an end by the *ALPAC Report* (1966) have surfaced again, like some Coelacanth from the deep, long believed extinct. Such systems were sold for many years under such trade names as LOGOS, XYZYX, SMART, Weidner and SYSTRAN; and the last, and best known, has been used for thirty years by the EU in Paris and Luxembourg.
- ※ **Secondly**, some large-scale, more theoretically based, MT projects continued, usually based in Universities, and have been tested in use, though sometimes on a scale smaller than that originally envisaged. METEO, for example, in Montreal, which was to have translated official documents from English to French, is still in use for the translation of the more limited world of weather reports.
- ⌘ **Thirdly**, workers in natural language in the field known as Artificial Intelligence (AI) began to make distinct claims about the need for their approach if there is ever to be general and high quality MT. Small pilot systems illustrating their claims were programmed, but their role in MT discussion was mainly of a theoretical nature.
- ☆ The **fourth** is certainly the revival of empirical statistical methods in MT, which began around 1989 and lost momentum in the 90s when the early systems, like Jelinek's failed to beat SYSTRAN decisively. However, empirical methods then colonised the whole of NLP, area by area, and now in the new millennium have returned to tackle MT itself.
- ❄ A possible **fifth** is that of hybrid methods, where researchers are seeking combinations of empirical methods with intelligent revivals of, earlier conceptual AI approaches.

✍ **2.** It is interesting to notice that the reactions of ***Bar-Hillel*** and AI workers like ***Minsky*** were in part the same: ***Minsky*** (1968) argued that MT clearly required the formalization of human knowledge for a system that could be said to understand, or as ***Bar-Hillel*** reviewed the situation in 1971: "It is now almost generally agreed upon that high-quality MT is possible only when the text to be translated has been understood, in an appropriate sense, by the translating mechanism".

✍ **3.** *Figure 2.1* is a formal structure of semantic primitives expressing the meaning of the action "drink": that drinking is a CAUSing to MOVE, preferably done by an ANlmate SUBJect (=agent) and to a liquid (FLOW STUFF), TO a particular ANImate aperture (THRU PART), and INto the SELF (=the animate agent). For short we will write *Fig. 2.1* as [drink].

The text structures in this system are semantic templates (together with semantic ties between them): a template is a network of formulas, containing at least an agent, action and object formula. Thus the template for "The adder drinks water" will be written: [the+adder drinks water] for short where the whole of *Fig. 2.1* is in fact at the central (action) node of that structure:

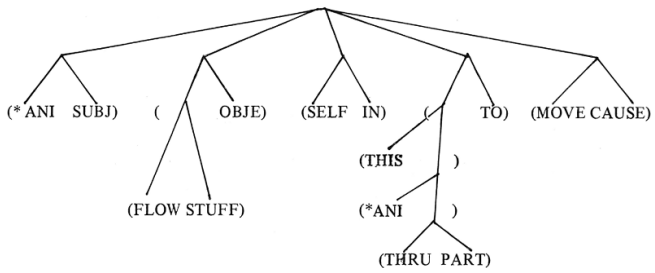
((*ANI SUBJ)(((FLOW STUFF)(OBJE)((*ANI IN)(((THIS(*ANI
(THRU PART)))TO)(BE CAUSE)))))

Fig. 2.1 A semantic formula for the action of drinking

4. The process of setting up the templates allows the formulas to compete to fill nodes in templates. Thus the formula for the (snake-) adder goes to the agent node in the template above in preference to the (machine-) adder because *Fig. 2.1* specifies, by (ANI SUBJ), that it prefers to be accompanied in a template by an animate agent formula. However, in the sentence: "*My car drinks gasoline*" the available formula for the first template node, namely [car], is not for an animate entity, yet it is accepted because there is no competitor for the position. An important later process is called extraction: additional template-like structures are inferred and added to the text representation even though they match nothing in the surface text. They are "deeper" inferences from the case structures of formulas in some actual template. Thus, to the template for [My car drinks gasoline] we would add an extraction (in double square parentheses in abbreviated form): [[gasoline in car]] which is an inference extracted from the containment subformula of *Fig. 2.1*, (SELF IN). Analogous extractions could be made for each case primitive in each formula in the template for [my car drinks gasoline]. After the programmed version of the system, reported in (Wilks 1978), a structural change (Wilks 1976b) allowed a wider, and more specific, form of expression in formulas by allowing thesaurus items, as well as primitives, to function in them. No problems are introduced by doing this, provided that the thesaurus items are also themselves words in the dictionary, and so have their formulas defined elsewhere in their turn. One advantage of this extension is to impose a thesaurus structure on the whole vocabulary, and so render its semantic expression more consistent.

5. *Fig. 2.2* The action formula for drinking installed at the central action node of a semantic template of formulas for "John drinks beer"

Here is the tree structure for the action of drinking:



6. In a template for [John drinks gin] the formula [drinks] (*Fig. 2.1 above*) shows within its tree that drinking is normally done by ***animate*** beings. So in [John drinks gin] the ***animate agent*** "preference" of [drinks] is satisfied by the presence of [John] (which can be seen to be animate because its head is MAN) at the agent node of the template that has [drinks] at its action node. The general preference rule of inference in the system is to take, as the provisional semantic representation at every stage, the template with the most satisfied preferences between its constituent formulas.

Chapter 3

An Artificial Intelligence Approach to Machine Translation

1. The aim of the **text processing** sections of the overall program is to derive from an English text an **interlingual representation** that has an adequate, though not excessive, complexity for two tasks:

- as a representation from which output in another natural language can be computed
- as a representation that can also serve as an analysandum of predicate calculus statements about some particular universe.

2. A fragmented text is to be represented by an **interlingual** structure consisting of TEMPLATES bound together by PARAPLATES and CS (or commonsense) INFERENCES. These three items consist of **FORMULAS** (and predicates and functions ranging over them and sub-formulas), which in turn consist of ELEMENTS. ELEMENTS are sixty primitive semantic units used to express the semantic entities, states, qualities, and actions about which humans speak and write. The elements fall into five classes as follows (elements in upper case):

- **entities:** MAN(human being), STUFF(substances), THING(physical object), PART(parts of things), FOLK(human groups), ACT(acts), STATE(states of existence), BEAST(animals), etc.
- **actions:** FORCE(compels), CAUSE(causes to happen), FLOW(moving as liquids do), PICK(choosing), BE(exists) etc.

- **type indicators:** KIND(being a quality). HOW(being a type of action) etc.
- **sorts:** CONT(being a container), GOOD(being morally acceptable), THRU(being an aperture) etc.
- **cases:** TO(direction). SOUR(source), GOAL(goal or end), LOCA(location), SUBJ(actor or agent), OBJE(patient of action), IN(containment), POSS(possessed by) etc.

3. FORMULAS are constructed from elements and right and left brackets. They express the senses of English words with one formula to each sense. The **formulas** are binarily bracketed lists of whatever depth is necessary to express the word sense. They are written and interpreted with – in each pair at whatever level it comes – a dependence of left side on corresponding right. **Formulas** can be thought of, and written out, as binary trees of semantic primitives. In that form they are not unlike the lexical decomposition trees of Lakoff and McCawley. Consider the action "drink" and its relation to the formula:

((*ANI SUBJ)(((FLOW STUFF)OBJE)((*ANI IN)(((THIS(*ANI (THRU PART)))TO)(BE CAUSE)))))

4. *ANI here is simply the name of a class of elements, those expressing animate entities namely, MAN, BEAST, and FOLK (human groups). In order to keep a small usable list of semantic elements, and to avoid arbitrary extensions of the list, many notions are coded by conventional sub-formulas: so, for example, (FLOW STUFF) is used to indicate liquids and (THRU PART) is used to indicate apertures.

5. A **formula** expresses the meaning of the word senses to which it is attached. This claim assumes a common sense distinction between explaining the meaning of a word and knowing facts about the thing the word indicates. The **formulas** are intended only to express the former, and to express what we might find – though in a formal manner – in a reasonable dictionary.

6. *Fig. 3.2* Correspondence of template head triples to sentence words:

i. John ran quickly.

MAN	MOVE	DTHIS
John	ran	[DTHIS]
	↑	
	quickly	

ii. John hit Bill.

MAN	DO	MAN
John	hit	Bill

iii. John gave Bill a ball.

MAN	GIVE	THING
John	gave	ball
	↑	↑
	(to) Bill	a

iv. John is in the house.

MAN	BE	DTHIS		DTHIS	PBE	THING
John	is	[DTHIS]		[DTHIS]	in	house
						↑
						the

v. John is sick
MAN BE KIND
John ↔ is ↔ boy
↑
a

iv. John is a boy
MAN BE MAN
John ↔ is ↔ father
↑
my

✍**7.** The **PICKUP** routines match bare templates onto the string of formulas for a text fragment. 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 so to reduce the number of templates matching a fragment, down to one if possible.

✍**8.** The **role of EXTEND** in general terms: it inspects the strings of formulas that replace a fragment and seeks to set up dependencies of formulas upon each other. It keeps a score as it does so, and in the end it 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:

⧈ those between formulas whose heads are part of the bare template

⌘ those of formulas whose heads are not in the bare template upon those formulas whose heads are in the bare template.

✍**9.** *Fig. 3.4* **Dependencies** between sentence words, where the upper line are words corresponding to template formula heads, including a dummy (DTHIS):

John ⟷ talked ⟷ [DTHIS]
↑
quickly

✍**10.** The **subtypes of dependence** are as follows:

A. among the formulas whose heads constitute the bare template

i. preferred subjects on actions
"John talked".

ii. preferred objects of actions on actions
"interrogated a prisoner.

B. of formulas not constituting bare templates on those that do,

i. qualifiers of substantives on substantive
"red door"
ii. qualifiers of actions on actions
"opened quickly"
iii. articles on substantives
"a book"
iv. of – fo phrases on substantives
"the house of my father fo"
v. qualifiers of actions on qualifiers of substantives
"very much"
vi. post verbs on actions
"give up"
vii. indirect objects on actions
"gave John a . . . ".
viii. auxiliaries on actions
"was going"
ix. "to" on infinitive form of action.
"to relax".

11. An **MT program** has to get "*Je bois du vin*" for " I drink wine" but to "*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:

❶ *THIS *DO *ANY where *THIS is any substantive head (not DTHIS) *DO is any real action head (not BE, PDO, DBE) *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.

1a. type *THIS BE KIND is treated with type 1.

❷ 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.

❸ *THIS DBE DTHIS Templates of this type represent the subject, split off from its action represented by type 2 template above The translation is simply generated from the stereotype of the subject formula, since the rest are 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).

❹ DTHIS PDO *REAL Templates of this type represent prepositional 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).

✍**12.** The general strategy for the final stages of the **MT 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.

Chapter 4

It Works but How Far Can It Go:
Evaluating the SYSTRAN MT System

✍**1.** I do not wish to suggest that the only challenge to SYSTRAN in MT comes from the use of **statistical techniques**. On the contrary, a number of researchers in linguistics and artificial intelligence (AI) (e.g. Sergei Nirenburg) continue to claim that advances in **linguistics**, **semantics** and **knowledge representation** during the past decades now permit the construction of a wholly new MT system that will be able to break through the quality ceiling of about 65–70% correctly-translated sentences established by **SYSTRAN**.

✍**2.** While **SYSTRAN** translates to a reasonably high degree of proficiency, it has no underlying theory that a theoretical linguist would acknowledge as such. **SYSTRAN** has been the subject of few published descriptions most of which have described it in terms of multiple passes through texts and the extraction of phrases and clauses by a "partial parser". But, in fact, there is good reason to believe that **SYSTRAN'S** original Russian-English performance is as good as it is because of the very large number of long word-collocations in Russian

✍**3.** The arguments for monolingual evaluation in the BR (**Battelle Report**) survey of evaluation methods were twofold:

✪ first, that estimates of the **fidelity** (correctness) of a translation strongly correlate to estimates of its quality in monolingual judgments of the output.

※ secondly, that a monolingual expert can be expected to judge the **overall coherence** of an output text since, as a text lengthens, the chances of its being both coherent and incorrect approach zero.

✍**4. BR** counted as distinct the following three concepts that they completely failed to distinguish, namely: **intelligibility**, **comprehensibility** and **readability**. At first glance, it might seem that the difference between

these categories is one of scale (with only comprehensibility applying to entire texts), but the intelligibility test is also applied by them to long sequences of output. Likewise, readability which "measures the appropriate overall contextual cohesiveness" of a text has little obvious contrast to the previous two categories. Indeed, the three separate tests given (one rating output on a "**clarity scale**", the second asking questions about the content of the output, and the third "**Cloze technique**" requiring a subject to fill in word gaps left at regular intervals in the output) could be applied equally to any of the three concepts with no change in the results. What is actually being discussed here are three different methods of measuring **coherence**, nothing more.

✍**5.** While the **Battelle Report** was poorly argued and statistically flawed, it provided us with the methodological basis for a new study.

✍**6.** The most significant finding was the 20% **carry-over effect** from updated to control text (balance of improvement: 30% of sentences improved minus 10% worsened) in a very different subject area from the one for which the system was originally developed.

✍**7.** There was a very **high variance** among **evaluators**, especially monolinguals. This was reduced to a significant result by distinguishing between sentences deemed judgeable and, of those judgeable, taking those deemed improved. While variance as to what was judgeable remained high, in the vital category of which judgeables were improved, variance was minimal: a strong, indirect confirmation of our methodology.

✍**8.** Since the question of **naturalness** of English output produced an odd response in the bilinguals, it is better ignored, especially since this notion is of little importance to the ultimate monolingual user, in any case.

PART III
MT Present

Chapter 5
Where Am I Coming From:
The Reversibility of Analysis and Generation in Natural Language Processing

✍**1.** ***Chomsky***'s original transformational-generative (**TG**) grammar project (1957) 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*.

✍ **2. Semantic parsing** 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 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.

Chapter 6

What are Interlinguas for MT:

Natural Languages, Logics or Arbitrary Notations?

✍ **1. SYSTRAN** was **not** an **interlingual system** because its power came largely from its bilingual dictionaries, and, as a matter of definition, a bilingual dictionary is language-pair dependent, and therefore a transfer, device. At that level of strictness, there have been very few truly interlingual MT systems, (i.e. without a bilingual dictionary). Probably the only historical candidate is Schank's MARGIE system of 1972, which did some English-German MT via a conceptual dependency (CD) representation.

✍ **2.** It is widely believed that **NLs** have their ambiguities resolved in use, up to some acceptable level, and that extensions of sense take place all the time, whether rule governed (e.g. as in Pustejovsky's generative lexicon (1995)), or, as in the old **AI/NLP** tradition, by means of manipulations on lexicons and knowledge structures that were general procedures but not necessarily equivalent to lexical rules. What would it be like, and I have no clear answer, to determine that the primitives of an **IL** (intermediate language) representation were in this position, too? Schank did, after all split the early TRANS into MTRANS, ATRANS and then others, so the suggestion has precedent.

Chapter 7

Stone Soup and the French Room:

The Statistical Approach to MT at IBM

✍ **1.** We need to establish a ground zero on what the **IBM system** is: their rhetorical claim is (or perhaps was) that they are a pure statistical system,

different from their competitors, glorying in the fact that they did not even need French speakers. By analogy with Searle's Chinese Room (Searle, 1980), one could call this theirs a French Room position: MT without a glimmering of understanding or even knowing that French was the language they were working on!

✍**2.** In essence, the method is an adaptation of one that worked well for **speech decoding**. The method establishes two components: **(a)** a trigram model of the sequences in the target language; **(b)** a model of quantitative correspondence of the parts of aligned sentences between French and English. The first is established from very large monolingual corpora in the language, of the order of 100 million words, the second from a corpus of aligned sentences in a parallel French-English corpus that are translations of each other.

✍**3.** In one sense, what **IBM** have done is partially automate the **SYSTRAN** construction process: replacing laborious error feedback with statistical surveys and lexicon construction. The **problem IBM** have is that few such vast bilingual corpora are available in languages for which MT is needed.

✍**4.** The basic **AI** argument for knowledge-based processing does not admit defeat and retreat, it just regroups. It has to accept **Bar Hillel**'s old anti-MT argument (Bar-Hillel 1960) on its own side – i.e. that as he said, good MT must in the end need knowledge representations. One version of this argument is the primitive psychological one: **humans** do **not** do translation by exposure to such vast texts, because they simply have not had such exposure, and in the end how people do things will prove important. Note that this argument makes an empirical claim about human exposure to text that might be hard to substantiate.

Chapter 8

The Revival of US Government MT Research in 1990

✍**1. Machine translation** remains the paradigm task for natural language processing. Unless **NLP** can succeed with the central task of machine translation, it cannot be considered successful as a field. We maintain that the most profitable approach to MT at the present time (1990) is an **interlingual** and **modular** one.

✍**2. MT** is one of the few computational tasks falling broadly within artificial intelligence (AI) that combine a fundamental intellectual research challenge with enormous proven need.

✍**3.** The vulgarized version of the history of **MT** is as follows: In the 1950s and 1960s large funds were made available to US MT which proved to

be a failure. The **ALPAC report** (1966) said MT was impossible and doomed all further US funding. **MT** work then moved to Canada and Europe where it partly succeeded, which was then followed by highly successful exploitation in Japan. The **truth**, of course, is **not** at all like that.

✍ **4.** MT work did **not stop** in the US after **ALPAC**: the AFOSR continued to fund it in the US and there were and are enormous commercial developments subsequently (the best known commercial systems being SYSTRAN, ALPS, LOGOS, METAL and SMART).

✍ **5. ALPAC** did **not** say MT was impossible nor that the work done was no good: only that at that point in history, with the cost and power of 1960s computers, human translation was cheaper.

✍ **6. MT** work did **not** move to Europe, since much of it stopped there also in response to the ALPAC report. The UK believed the **ALPAC report**, and only in France did serious work continue, where the GETA system in Grenoble became the foundation for a range of others, including the major Japanese university system (Mu) and aspects of the Eurotra system, which was designed to be a multilingual system between the languages of the EEC.

✍ **7.** One way in which all **MT** work is in **SYSTRAN's debt** is that it is the main existence of MT proof: it convinces doubters that there that machine translation now exists, albeit in primitive form, and can be purchased on a large scale and at a quality that many users find acceptable for their needs. A key defect in the **ALPAC report** was that it underestimated how large a market there was for partially accurate, low quality, MT, and SYSTRAN filled that market. The point now, of course, is to move on to the huge market for higher-quality MT.

✍ **8.** Steady developments in various aspects of **NLP** make available large portions of an MT system more or less off the shelf, which greatly facilitates the construction of new MT systems. These developments are the following:

- ☆ **Clearer understanding of semantics**: Recent refinements of taxonomical ontologies of representation provide an interlingua-like basis for a new, more powerful, MT. Making maximal use of the high-level linguistic and semantic generalizations shared among languages, one can minimize or even eliminate language-tolanguage lexical or structural transfer rules and so increase the portability of the system across domains.
- ❄ **More complete grammars**: Development of grammars is an ongoing process. There exist today grammars that cover English (and other languages such as German, Chinese, Japanese, and French) far more

extensively than the most comprehensive grammars of 20 years ago did.

- **Better existing generation and parsing technology**: Single-sentence parsing and generation has been studied to the point where a number of well-established paradigms and algorithms exist, each with known strengths and weaknesses, a situation which greatly facilitates the construction of a new MT system (in fact, in recent years a number of general-purpose generators have been distributed: Penman, Mumble, Frege, etc.).

9. Statistics, although not a complete translation paradigm, plays several important roles in MT, however such as selecting the most normative (frequent) rendition into words in each target language. Statistics can select collocations from large text corpora (such as the preferred use of "pitch black" rather than "asphalt black"). Given a large potential lexicon, simple frequency analysis can direct the dictionary-building work towards the most frequent words first, so as to obtain maximal utility of a system during development phases. All evaluation metrics of fluency, accuracy and cost of translation are **statistically** based.

10. Modularity is independent of interlinguality though opting for the latter requires the former. Strong modularity of language components would now be supported by most researchers and developers in MT, largely because it allows the addition of new languages with minimum dislocation. The advantages of this **modular approach** include the following:

- Various projects and various theoretical approaches will be able to participate.
- Projects need not have experience in all aspects of MT to participate.
- Redundant development of modules will be eliminated.
- Interproject collaboration will be stimulated throughout the U.S.

11. In order to produce MT of superior quality that existing systems, one of the most powerful key ideas is the use of **discourse-related** and pragmatic terms. Most MT systems operate on a *sentence-by-sentence* basis only; they take no account of the **discourse structure**. Given recent work on discourse structure at various centers in the U.S., structural information should be taken into account and can be used to improve the quality of the translation. Similarly, **pragmatic information**, such as Speech Acts, reference treatment, and perhaps even some stylistic notions (to the extent that notations have been developed to represent them) will be used to improve the quality of the translation.

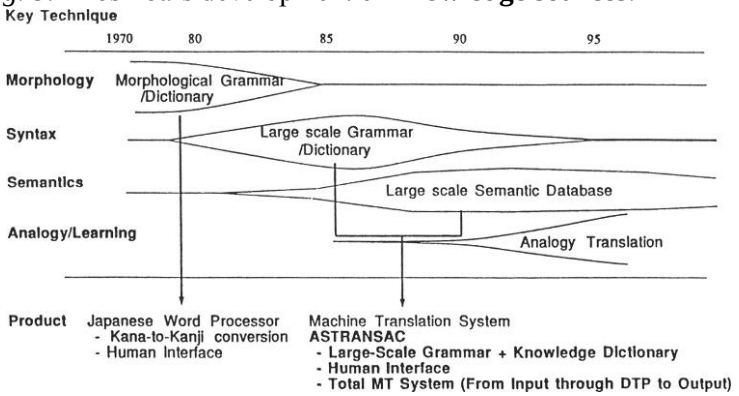
Chapter 9

The Role of Linguistic Knowledge Resources in MT

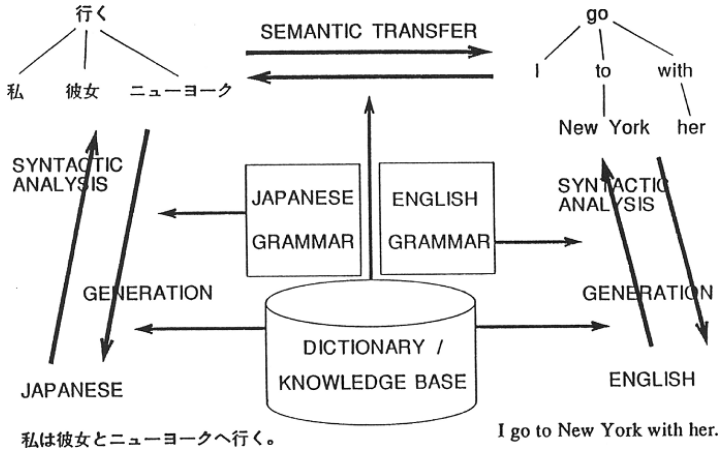
1. Most MT systems make use of at least some, or possibly all of the following kinds of **lexical knowledge sources** as distinct from corpora alone (as in Brown et al.):

- Morphology tables
- Grammar rules
- Lexicons

2. *Fig. 9.1* Toshiba's development of **knowledge sources**:



3. In the figure below (*Fig. 9.2*), Toshiba indicated their system's overall translation procedure. Without committing ourselves to a specific view of what "**semantic transfer**" means, we can infer that the bolder arrows represent the translation tasks to be performed, while the lighter arrows indicate Toshiba's view of where the knowledge forms they emphasize (a merger of our latter two items, omitting morphology) distribute across those tasks.

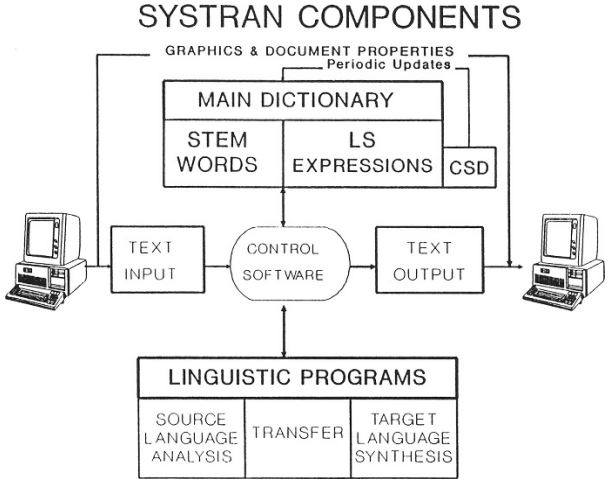


4. SYSTRAN has been described (at least in parody) as utilizing no **knowledge sources**; it has been thought of by some as having, in effect,

a mere sentence dictionary of source and target languages. Nor is this notion as absurd as linguists used to think: the number of English sentences under fifteen words long, for instance, is very large, but not infinite.

✍**5.** So, on the above definition, an MT system that did MT by such a method of *direct one-to-one* sentence pairing would definitely not have a **knowledge source**. But, although part of the success of the Dayton-based SYSTRAN Russian-English system is certainly due to its roughly 350K lexicon of phrases, idioms, and semi-sentences (Wilks, 1991), **SYSTRAN** does **not** really conform to this parody of it (Toma, 1976). It is interesting to note in passing that, utterly different as they are on a symbolic-statistical spectrum, SYSTRAN and CANDIDE have earned similar opprobrium from linguists!

✍**6.** *Fig. 9.6* The use of **dictionaries** in SYSTRAN:



✍**7. SYSTRAN** is a strongly **lexically-dependent** MT system, and its JE and EJ modules were owned in Japan at the time of writing (by Iona) and are therefore technically speaking Japanese systems. SYSTRAN's JE and EJ modules have three types of dictionaries, and are described in the company's own words as follows (SYSTRAN, 1991):

- A "**word boundary**" dictionary for matching words and establishing word boundaries in Japanese text, where each word is not clearly bounded by spaces (as in English and other European languages).
- A "**stem**" dictionary containing source language words and their most frequently used target language equivalents. This dictionary also contains morphological, syntactic, and semantic information about each entry word.

✪ A "**limited semantics**" (LS) dictionary of expressions, special collocations, and macro instructions for handling translation problems of low to medium complexity.

Chapter 10

The Automatic Acquisition of Lexicons for an MT System

1. ULTRA (Universal Language TRAnslator) is a multilingual, interlingual machine translation system which currently translates between five languages (Chinese, English, German, Japanese, Spanish) with vocabularies in each language based on about 10,000 word senses. It makes use of recent **AI**, **linguistic** and **logic programming** techniques, and the system's major design criteria are that it be robust and general in purpose, with simple-to-use utilities for customization. Its **special features** include:

- ※ a multilingual system with a language-independent system of intermediate representations (interlingual representations) for representing expressions as elements of linguistic acts;
- ☆ bidirectional Prolog grammars for each language incorporating semantic and pragmatic constraints;
- ❄ use of relaxation techniques to provide robustness by giving preferable or "near miss" translations;
- ❋ access to large machine-readable dictionaries to give rapid scaling up of size and coverage;
- ⌖ multilingual text editing within Xwindows interface for easy interaction and document preparation in specific domains (e.g., business letters, proforma memoranda, telexes, parts orders).

2. The **interlingual representation** (IR) has been designed to reflect our assumption that what is universal about language is that it is used to perform acts of communication: asking questions, describing the world, expressing one's thoughts, getting people to do things, warning them not to do things, promising that things will get done and so on.

3. Translation, then, can be viewed as the use of the target language to perform the same act as that which was performed using the source language. The IR serves as the basis for analyzing or for generating expressions as elements of such acts in each of the languages in the translation system.

4. There are two types of entries related to the specification of a *lexical item* in the **ULTRA** system:

- ❁ those for intermediate representation (IR) word sense tokens
- 🗎 those for the words of the individual languages

✍**5.** The **ULTRA system**, designed at built at NMSU, was the second practical system in which I was involved as which is reported in this book: the first was the *Stanford semantics-based system*, built by two people (Annette Herskovits and myself), using the system I had designed for my thesis work. The second was this **ULTRA system**, funded in part by New York investors but never developed commercially by them. It was never properly evaluated and its only interesting feature may be that it used much of the architecture of the **EUROTRA** system and finally achieved better performance that that system at perhaps 1% of the cost. The third system was the multi-site **PANGLOSS** system, funded by **DARPA** in competition with IBMs **CANDIDE**, which proved in the end very much the way forward at that time, even though **PANGLOSS** led to usable machine-aided spin offs and made points about the sue of resources, linguistics and semantics that have not disappeared altogether and will be discussed the final chapters.

PART III
MT Future

Chapter 11
Senses and Texts

✍**1.** Yarowsky contrasts his work with that of efforts like (Cowie et al. 1992) that were **dictionary based**, as opposed to (unannotated) **corpus based** like his own. But a difference he does not bring out is that the Cowie et al. work, when optimized with simulated annealing, did go through substantial sentences, mini-texts if you will, and sense-tag all the words in them against LDOCE at about the 80% level. It is not clear that doing that is less useful than procedures like Yarowsky's that achieve higher levels of **sense-tagging** but only for carefully selected pairs of words, whose sense-distinctions are not clearly dictionary based, and which would require enormous prior computations to set up ad hoc sense oppositions for a useful number of words.

✍**2.** These are still early days, and the techniques now in play have probably not yet been combined or otherwise optimised to give the best results. It may not be necessary yet to oppose, as one now standardly does in MT, large-scale, less accurate, methods, though useful, with other higher-performance methods that cannot be used for practical applications. That the field of **sense-tagging** is still open to further development follows if one accepts the aim of this chapter which is to attack two claims, both of which are widely believed, though not at once: that **sense-tagging** of

corpora cannot be done, and that it has been solved. As many will remember, MT lived with both these, ultimately misleading, claims for many years.

Chapter 12
Sense Projection

1. What all **AI projects**, of whatever level, have in common is an appeal to very general knowledge and principles, coupled to the claim that **MT** work must take account of these if it is ever to achieve *generality* and *reliability*. The reply to this claim, from experience with projects like **SYSTRAN**, is that the **AI** examples that make these points are artificial and/or rare, and they can be ignored for practical purposes. This is clearly an empirical dispute and open to test.

Chapter 13
Lexical Tuning

1. Automatic **word-sense disambiguation** (WSD) is now an established modular task within empirically-based computational linguistics and has been approached by a range of methods sometimes used in combination. These experiments are already showing success rates at, or close to, the target ninety-five-per-cent levels attained by established modules like part of speech tagging in the mid-Nineties.

2. WSD is different in a key respect from tasks like part-of-speech tagging (**POS**): namely, that lexicons need to adapt dynamically in the face of new corpus input.

3. The contrast here is in fact quite subtle, as can be seen from the interesting intermediate case of semantic tagging: attaching semantic, rather than POS, tags to words automatically, a task which can then be used to do more of the **WSD** task than **POS tagging** can, since the ANIMAL or BIRD versus MACHINE tags can then separate the main senses of "crane". In this case, as with **POS**, one need not assume any novelty in the tag set, in the sense of finding in the middle of the task that one needs additional tags. But one must also allow for novel assignments from the tag set to corpus words, for example, when a word like "dog" or "pig" was first used in a human sense. It is just this sense of novelty that **POS tagging** also has, of course, since a **POS tag** like VERB can be applied to what was once only a noun, like "ticket". This kind of assignment novelty, in **POS** and **semantic tagging**, can be premarked up with a fixed tag inventory, hence both these techniques differ from

genuine sense novelty which, we shall argue, cannot be premarked in any simple way.

4. This latter aspect, which we shall call **Lexical Tuning**, can take a number of forms, including:

- adding a sense to the lexical entry for a word
- adding an entry for a word not already in the lexicon
- adding a subcategorization or preference pattern etc. to an existing sense entry

Chapter 14

What Would Pragmatics-Based Machine Translation be Like?

1. These are tasks that people perform and which they use language in performing. Following Morgan (Morgan, TINLAP-2, 109), we have analyzed such tasks as having three general aspects: **referential information** content constituting what is said, **stylistic information** content constituting how it is said, **communicative information** content constituting why it is said, and an intermediate representation (IR) has been explicitly designed to represent all three aspects without regard to any particular language.

2. Translation, then, can be viewed as the use of the target language to perform the same act as that which has been performed using the source language and the IR serves as the basis for analyzing or for generating expressions as elements of such acts in each of the languages in the translation system

3. The two **CRL approaches** are all extensions of fairly well established lines of research and are consistent with a certain position on computational semantics. The main **assumptions** of this position are two-fold:

- First, the **problem of the word sense is inescapable**: lexical ambiguity is pervasive in most forms of language text, including dictionary definitions, hence the words used in dictionary definitions of words and their senses are themselves lexically ambiguous and must be disambiguated.
- Second, **knowledge and language are inseparable**, i.e., that the semantic structure of language text and of knowledge representations share common organizing principles, and that some kinds of language text structures are a model for knowledge structures (Wilks 1978).

4. The main differences between the **three CRL approaches** are over what we call **bootstrapping**, i.e., over what knowledge, if any, needs to

be hand-coded into an initial analysis program for extracting semantic information from a MRD, and the kinds of knowledge they produce for a MTD. Both the approaches begin with a degree of hand-coding of initial information but are largely automatic. In each case, moreover, the degree of hand-coding is related to the source and nature of semantic information sought by the approach.

- **Approach I**, a statistically based approach, uses the least hand-coding but then the co-occurrence data it generates is the simplest form of semantic information produced by any of the approaches.
- **Approach II** requires the hand-coding of a grammar and semantic patterns used by its parser, but not the hand-coding of any lexical material. This is because the approach builds up lexical material from sources wholly within LDOCE (*Longman's Dictionary of Contemporary English*).

5. The **ultimate goal** of *machine translation* is undoubtedly to implement a program on an architecture which takes as **input** an utterance (in the form of text or speech) in one **natural** language and produce as **output** an equivalent utterance in a second (distinct) **natural** language. This is clearly a problem of software and hardware engineering. At one level, then, the goal of theoretical MT should be to identify the inherently limiting computational characteristics of the proposed approaches and provide alternative approaches which overcome those limitations. However, identifying such limitations and working out alternative approaches is dependent on assumptions about natural languages, and natural language processing in general, such as what the nature of natural language is, what the nature of natural language understanding is, what the nature of natural language production is, and, in the case of machine translation, what the nature of natural language translation is.

Chapter 15

Where was MT at the End of the Century:
What Works and What Doesn't?

1. What was not foreseen in the 1990s was the way that the **Web/Internet** would transform the resource issue by simply becoming THE monolingual language resource, at least for alphabetic languages, so that researchers began to talk of billions of words in corpora and The Web as Corpus. Sharing has much improved with the entrenchment of the resource associations mentioned above and the growth of additional competitions tied to annotated resources for specific linguistic functions: SENSEVAL, PARSEVAL etc. These have now grown out of the direct

control of the US military authorities to civil society and to groups associated with language families e.g. ROMANSEVAL.

✍ **2.** But even now it is difficult to get resources for **NP** areas like dialogue processing because both the speech recognition and **NLP** processing in dialogue tends to be tied tightly to application domains and so classic existing resources do not always help researchers much if they are not in the right domain. The availability of parallel language resources, specifically for **MT**, has improved with the availability of corpora from international banks and the documentation of the **EU** itself, but the greatest source of resources for quantitative **MT** (see next chapter) has been the growing data provided by translation bureau themselves.

Chapter 16
The Future of MT in the New Millennium

✍ **1.** There has certainly been a change in techniques: most importantly the use of the **BLEU** technique and its variants (**NIST**, **ROUGE**, **METEOR** etc.) have virtually replaced all the earlier techniques mentioned earlier such as **Cloze tests** etc. Their advantage is seen as being lessened of human involvement while giving clear quantitative results. Basically, all such tests rest on comparison of ngrams (up to trigrams) between the translations to be rated and a canonical translation.

✍ **2.** So, where will more advanced **MT** come from? I take it for granted that advance in **MT** will come from "**phenomena of scale**": the use of very large dictionaries in particular, and the extraction from them, and from large text samples, of collocational, semantic and pragmatic information, as well as new techniques for combining these sources in differing circumstances. It will also require, as it has with historical **MT**, an understanding of, as well as techniques for, maintaining and adapting very large programs whose original structure has become obscure.

Short Answer Items & Tests

3.2 Short Answer Items

1. BR counted as distinct the following three concepts that they completely failed to distinguish, namely: intelligibility, ……. and readability.

2. In one sense, what IBM have done is partially automate the SYSTRAN construction process: replacing laborious error feedback with ……. surveys and ……. construction. The problem IBM have is that few such vast bilingual corpora are available in languages for which MT is needed.

3. Most MT systems make use of at least some, or possibly all of the following kinds of lexical knowledge sources as distinct from corpora alone: ……. tables, ……. rules, Lexicons.

4. SYSTRAN is a strongly …….-dependent MT system, and its JE and EJ modules were owned in Japan at the time of writing (by Iona) and are therefore technically speaking Japanese systems.

5. Following Morgan, we have analyzed such tasks as having three general aspects: ……. information content constituting what is said, ……. information content constituting how it is said, ……. information content constituting why it is said, and an intermediate representation (IR) has been explicitly designed to represent all three aspects without regard to any particular language.

3.3 Answers

1) comprehensibility	**2)** statistical, lexicon
3) Morphology, Grammar	**4)** lexically
5) referential, stylistic, communicative	

☙☆❧ 3.4 Tests ☙☆❧

✍ Select the best choice.

1. The arguments for monolingual evaluation in the BR (Battelle Report) survey of evaluation methods were twofold: first, that estimates of the ……. (or …….) of a translation strongly correlate to estimates of its quality in monolingual judgments of the output. Secondly, that a monolingual expert can be expected to judge the overall ……. of an output text since, as a text lengthens, the chances of its being both ……. and ……. approach zero.

a) fidelity, accuracy, clarity, incoherent, inaccurate
b) fidelity, correctness, coherence, coherent, incorrect
c) correctness, coherence, cohesion, cohesive, accurate
d) accuracy, correctness, cohesion, coherent, accurate

2. Semantic parsing is a method claiming that text can be parsed to an appropriate representation without the use of an explicit and separate ……. component.

a) lexical
b) lexical and grammatical
c) grammatical or stylistic
d) syntactic

3. According to the vulgarized version of the history of MT, in the 1950s and 1960s large funds were made available to US MT which proved to be a ……. . The ……. report (1966) said MT was ……. and ……. all further US funding.

a) success, Bar Hillel, promising, encouraged
b) failure, Bar Hillel, impossible, doomed
c) failure, ALPAC, impossible, doomed
d) success, ALPAC, promising, encouraged

4. The main differences between the three CRL approaches are over what we call ……., i.e., over what knowledge, if any, needs to be hand-coded into an initial analysis program for extracting semantic information from a MRD, and the kinds of knowledge they produce for a MTD.

a) interlingua encoding
b) bootstrapping
c) lexical representation
d) syntactic parsing

5. Which item is correct? The ALPAC report (1966) ……. .

a) did not say MT was impossible
b) did not say that the work done was no good

c) said that at that point in history, with the cost and power of 1960s computers, human translation was cheaper
d) all of the above are correct

☙✺❧ 3.5 Answer key ☙✺❧

	a	b	c	d		a	b	c	d
1		×			2				×
3			×		4		×		
5				×					

Book ❹

Computers and Translation: A translator's guide

Harold Somers

4.1 Notes

Chapter 1
Introduction
Harold Somers

1. In **1964**, the US government decided to see if its money had been well spent, and set up the Automated Language Processing Advisory Committee (**ALPAC**). Their report, published in 1966, was highly negative about MT with very damaging consequences. Focusing on Russian–English MT in the USA, it concluded that MT was slower, less accurate and twice as expensive as human translation, for which there was in any case not a huge demand. It concluded, infamously, that there was "no immediate or predictable prospect of useful machine translation". In fact, the **ALPAC report** went on to propose instead fundamental research in computational linguistics, and suggested that machine-aided translation may be feasible. The damage was done however, and MT research declined quickly, not only in the USA but elsewhere.

2. The 1970s and early 1980s saw MT research taking place largely outside the USA and USSR: in Canada, western Europe and Japan, political and cultural needs were quite different. **Canada's** bilingual policy led to the establishment of a significant research group at the University of Montreal.

3. In **Europe** groups in France, Germany and Italy worked on MT, and the decision of the Commission of the European Communities in Luxembourg to experiment with the Systran system (an American system which had survived the **ALPAC** purge thanks to private funding) was highly significant.

✍**4.** In **Japan**, some success with getting computers to handle the complex writing system of Japanese had encouraged university and industrial research groups to investigate Japanese–English translation.

✍**5.** By the mid **1980s**, it was generally recognized that fully automatic high-quality translation of unrestricted texts (**FAHQT**) was not a goal that was going to be readily achievable in the near future. Researchers in MT started to look at ways in which usable and useful MT systems could be developed even if they fell short of this goal.

✍**6.** Many commentators now distinguish between the use of MT for **assimilation**, where the user is a reader of a text written in an unfamiliar language, and **dissemination**, where the user is the author of a text to be published in one or more languages.

✍**7.** In particular, the idea that MT could work if the input text was somehow restricted gained currency. This view developed as the **sublanguage** approach, where MT systems would be developed with some specific application in mind, in which the language used would be a subset of the "**full**" language, hence "**sublanguage**".

✍**8.** MT researchers continue to set themselves ambitious goals. **Spoken-language translation** (SLT) is one of these goals. **SLT** combines two extremely difficult computational tasks: speech understanding, and translation. The *first* task involves extracting from an acoustic signal the relevant bits of sound that can be interpreted as speech (that is, ignoring background noise as well as vocalizations that are not speech as such), correctly identifying the individual speech sounds (phonemes) and the words that they comprise and then filtering out distractions such as hesitations, repetitions, false starts, incomplete sentences and so on, to give a coherent text message. All this then has to be **translated**, a task quite different from that of translating written text, since often it is the content rather than the form of the message that is paramount. Furthermore, the constraints of **real-time processing** are a considerable additional burden.

Chapter 2

The translator's workstation

Harold Somers

✍**1.** A **corpus** is a collection of text, usually stored in a computer-readable format. The example database of a translation memory is an example of a corpus, with the particularly interesting property of being an aligned **parallel corpus**, by which is meant that it represents texts which are translations of each other ("**parallel**"), and, crucially, the corpus has been

subdivided into smaller fragments which correspond to each other (hence "**aligned**").

✍**2.** The translator's **workstation** represents the most *cost-effective* facility for the professional translator, particularly in large organisations. It makes available to the translator at one terminal (whether at an individual computer or as part of a company network) a range of integrated facilities: multilingual word processing, electronic transmission and receipt of documents, spelling and grammar checkers (and perhaps style checkers or drafting aids), publication software, terminology management, text concordancing software, access to local or remote term banks (or other resources), translation memory (for access to individual or corporate translations), and access to automatic translation software to give rough drafts.

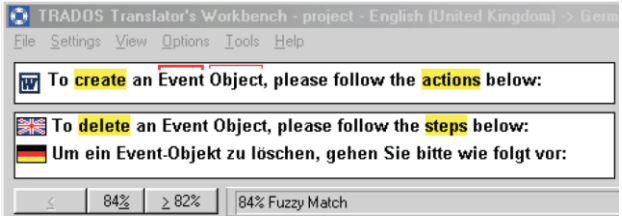
Chapter 3

Translation memory systems

Harold Somers

✍**1.** One of the most significant computer-based aids for translators is the now widely used **translation memory** (TM). First proposed in the 1970s, but not generally available until the mid 1990s, the idea is that the translator can consult a database of previous translations, usually on a sentence-by-sentence basis, looking for anything similar enough to the current sentence to be translated, and can then use the retrieved example as a model. If an exact match is found, it can be simply cut and pasted into the target text. Otherwise, the translator can use it as a suggestion for how the new sentence should be translated. The TM will **highlight** the parts of the example(s) that differ from the given sentence, but it is up to the translator to decide which parts of the target text need to be changed.

✍**2. Trados**'s translation memory window showing **partial match**:



✍**3.** The original idea for **TM** is usually attributed to **Martin Kay** who, as long ago as 1980, wrote a highly influential paper entitled "*The Proper Place of Men and Machines in Language Translation*" in which he proposed a basic blueprint for what we now call translator's workstations. In fact, the details relating to TMs are only hinted at obliquely: "… the translator

might start by issuing a command causing the system to display anything in the store that might be relevant to [the text to be translated] …. Before going on, he can examine past and future fragments of text that contain similar material." (Kay, 1980: 19)

4. A prerequisite for a TM system is of course a **database** of translation examples. Known to computational linguists as an "**aligned parallel corpus**", there are principally three ways of building a TM database:
- ⌘ building it up as you go along,
- ❁ importing it from elsewhere, or
- ⌖ creating it from a parallel text.

5. Perhaps the simplest method is to **build it up** as you go along. Each sentence you translate is added to the database. Obviously, if you are working on a text that is similar to one you worked on before, you can load up the database that you created last time and continue to add to it this time. Conversely, if you are working on different projects and want to develop separate databases for each of them, this can also be done. Unfortunately, this method of developing the database is painfully slow, and there will be a long lead time before the translator really feels the benefit of the software.

6. The next simplest method is to "**import**" the database from elsewhere. **TM** databases are not simply text files. In order for the matching algorithms to work efficiently, the databases have to be highly structured, with indexes to facilitate efficient retrieval of examples. Many **TM** systems also feature a terminology matching facility, or other add-ons. In particular, it is often the case that as items are added to the database they can be annotated with additional information such as their source, date, validation code, the name of the translator; and as they get used, some systems maintain statistics which can influence the matching algorithm so that it chooses more frequently used examples wherever possible. On top of this there is the question of compatibility of different word-processing formats. All of these elements and more are subject to **TMX** (*Translation Memory eXchange*) agreements.

7. The third, and technically most complex, alternative is to take an existing translation together with the original text and have the software build a **TM** database from it automatically. This involves **alignment** above all else, though as the previous paragraph indicated, once aligned there will be an amount of indexing and other database manipulations that need not concern us here.

8. Alignment involves matching up the source text and the translation segment by segment into translation pairs. "**Segments**" are usually understood to correspond to sentences or other more or less easily

distinguishable text portions, such as titles. If the translation is straightforward, then so is the alignment. But three factors can make alignment more difficult than it at first seems:

- one is the difficulty of accurately recognizing where sentences begin and end;
- the second is the fact that—depending on the language pair—a single sentence in one language may not necessarily correspond to a single sentence in the other language;
- the third factor is that translators may more or less freely change the relative order of sentences in the translation.

9. In "Example-based MT" (**EBMT**), **like** in **TM**s, there is an aligned parallel corpus of previous translations, and from this corpus are selected appropriate matches to the given input sentence.

10. In a **TM**, however, it is up to the user, the translator, to decide what to do with the retrieved matches.

11. In **EBMT**, we try to automate the process of selecting the best matches or fragments from the best matches, and then to "recombine" the corresponding target-language fragments to form the translation. Because this has to be done automatically by the system, any linguistic knowledge or translator's expertise that needs to be brought to bear on the decision has to be somehow incorporated into the system.

Chapter 4

Terminology tools for translators

Lynne Bowker

1. Terminology is the discipline concerned with the collection, processing, description and presentation of terms, which are lexical items belonging to specialized subject fields.

2. Effective **terminology** management can help to cut costs, improve linguistic quality, and reduce turn-around times for translation, which is very important in this age of intense time-to-market pressures.

3. Dating back to the 1960s, **term banks** were among the first linguistic applications of computers. Term banks are basically large-scale collections of electronic term records, which are entries that contain information about terms and the concepts they represent (e.g., definitions, contexts, synonyms, foreign language equivalents, grammatical information). Early **term banks** were originally developed by large corporations or institutions to serve as resources for in-house translators.

4. When desktop computers first became available in the 1980s, personal **TMS**s were among the first computer-aided translation (CAT) tools to be made commercially available to translators. Translators were able to use these tools to create and maintain personal **termbases**, in which they could record the results of their own terminological research.

5. One of the newest computer-aided terminology tools to arrive on the scene is the **term-extraction tool**. Essentially, this type of tool attempts to search through an electronic corpus and extract a list of candidate terms that a translator may wish to include in a **termbase**.

6. Both **term banks** and **termbases** are made up of data records called **term records**.

7. Term records treat a single concept and may contain a variety of linguistic and extralinguistic information associated with that concept in one or more languages. There are **no** *hard-and-fast* rules about what kind of information should be included on a term record—translators will have to decide this for themselves based on the availability of data and on the requirements of the project at hand. Nevertheless, types of information that may be found on **term records** could include: an indication of the subject field, equivalents in one or more languages, grammatical information (e.g., part of speech or gender), synonyms, definitions, contexts, usage notes (e.g., rare, archaic, British), and any other comments or information the translator thinks might be helpful in order to use the term in question correctly.

8. The most fundamental function of a **TMS** is that it acts as a repository for consolidating and storing terminological information for use in future translation projects. In the past, many TMSs stored information in structured text files, mapping source to target terminology using a unidirectional one-to-one correspondence. This caused difficulties, for example, if an English–French **termbase** needed to be used for a French–English translation. Contemporary **TMS**s tend to store information in a more **concept-based** way, which permits mapping in multiple language directions.

9. Once the terminology has been stored, translators need to be able to retrieve this information. A range of search and retrieval mechanisms is available. The simplest search technique consists of a simple look-up to retrieve an exact match. Some **TMS**s permit the use of wildcards for truncated searches. A wildcard is a character such as an asterisk (*) that can be used to represent any other character or string of characters. For instance, a wildcard search using the search string translat* could be used

to retrieve the term record for translator or the term record for translation, etc.

✍**10.** More sophisticated TMSs also employ **fuzzy matching** techniques. A fuzzy match will retrieve those term records that are similar to the requested search pattern, but which do not match it exactly.

✍**11. Fuzzy matching** allows translators to retrieve records for morphological variants (e.g., different forms of verbs, words with suffixes or prefixes), for spelling variants (or even spelling errors), and for multiword terms, even if the translators do not know the precise order of the elements in the multiword term. Some examples of term records that could be retrieved using fuzzy matching techniques are illustrated in the following Figure:

Search pattern entered by user:	*Term record retrieved using fuzzy matching:*
advertising	advertisement
organisation	organization
centre for preventing and controlling diseases	Center for Disease Control and Prevention

✍**12.** In cases where *wildcard searching* or **fuzzy matching** is used, it is possible that more than one record will be retrieved as a potential match. When this happens, translators are presented with a **hit list** of all the records in the **termbase** that may be of interest and they can select the record(s) they wish to view. The following Figure shows some sample **hit lists**:

*Hit list containing records that match the wildcard search pattern '*nut'*	*Hit list containing records that match the fuzzy search pattern 'post-office box number'*
coconut	Post Office
hazelnut	post office box
peanut	P. O. box number
walnut	postbox

✍**13.** The principal **advantages** of using a **TMS** rather than a card index: TMSs permit more *flexible* **storage** and **retrieval**. In addition, it is easier to update electronic information, and faster to search through electronic files. Even though a word processor allows information to be stored in electronic form, it is not an adequate tool for managing terminology in an efficient way, and its search facilities slow down considerably as the **termbase** grows in size.

✍**14. Term-extraction** tools can be either monolingual or bilingual. A **monolingual** tool attempts to analyze a text or corpus in order to identify candidate terms, while a **bilingual** tool analyzes existing source texts

along with their translations in an attempt to identify potential terms and their equivalents.

✍**15. Term-extraction** tools that use a linguistic approach typically attempt to identify word combinations that match particular part-of-speech patterns. For example, in English, many terms consist of adjective+noun or noun+noun combinations. In order to implement such an approach, each word in the corpus must first be associated with its appropriate part of speech, which can be done with the help of a piece of software known as a **tagger**. Once the corpus has been **tagged**, the term-extraction tool simply identifies all the occurrences that match the specified part-of-speech patterns.

✍**16. Term-extraction** tools that use a **statistical approach** basically look for repeated sequences of lexical items. The **frequency threshold**, which refers to the number of times that a sequence of words must be repeated, can often be specified by the user.

Chapter 5
Localisation and translation
Bert Esselink

✍**1. Localisation** is all about customising things (user manuals for products, especially software, and the products themselves) for a "**local**" audience.

✍**2.** The **Localisation** Industry Standards Association (LISA) defines localisation as follows: "**Localisation** involves taking a product and making it linguistically, technically, and culturally appropriate to the target locale where it will be used and sold."

✍**3.** Often, **localisation** is abbreviated as **L10n**, where **10** represents the number of letters between the **l** and **n**.

✍**4.** Making a product **linguistically** appropriate to a particular market basically means translating it, and making it technically appropriate means adjusting all product specifications to support standards in the target market. **Cultural adaptations** are modifications of the source text to reflect situations and examples common in the target market.

✍**5.** According to LISA, **internationalisation** is … the process of generalising a product so that it can handle multiple languages and cultural conventions without the need for re-design. **Internationalisation** takes place at the level of program design and document development.

✍**6.** LISA defines **globalisation** as follows: **Globalisation** addresses the business issues associated with taking a product global. In the **globalisation** of high-tech products this involves integrating *localization*

throughout a company, after proper *internationalisation* and product design, as well as marketing, sales, and support in the world market.

7. Differences between "**translation**" and "**localisation**" can categorised as follows:
- ✯ activities,
- ✠ complexity,
- ❄ adaptation level, and
- ❁ technology used.

8. Examples of activities in **localisation** that are ***not*** necessarily part of **traditional translation** are:
- ✷ multilingual project management,
- ✪ software and online help engineering and testing,
- ✵ conversion of translated documentation to other formats,
- ⌘ translation memory alignment and management,
- ✺ multilingual product support, and
- ⌖ translation strategy consulting.

9. CAT tools, also called machine-aided translation tools, can be categorised as follows:
- ◎ translation memory (TM) tools,
- ◈ terminology tools,
- ▣ software localisation tools.

10. When text that has been segmented by a **TM tool** is translated, all translations are automatically stored in the records containing the source segments. If **identical** or **similar** sentences occur in the source text, the translations are automatically retrieved from the database and inserted into the target text.

11. An identical segment that is automatically translated is called a **full match**; a similar sentence that is automatically translated is called a **fuzzy match**.

12. Obviously, **fuzzy matches** need to be **post-edited** to make them correspond to the source text. A **fuzzy match** is, for example, a sentence where only one word has changed compared to an already translated sentence.

13. In 1990, the Localisation Industry Standards Association, **LISA**, was founded in Switzerland. **LISA** defines its mission as "… promoting the localisation and internationalisation industry and providing a mechanism and services to enable companies to exchange and share information on the development of processes, tools, technologies and business models connected with localisation, internationalisation and related topics."

Chapter 6
Translation technologies and minority languages
Harold Somers

✍**1.** The **minority languages** are inferior in the provision of the whole range of computer aids for translators: not just MT systems, CAT systems, on-line dictionaries, thesauri, and so on, but even simple tools like spelling- and grammar-checkers.

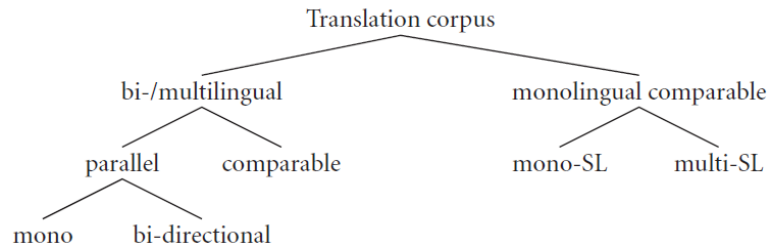
✍**2.** Optical character recognition (**OCR**) is a process that converts scanned images into text. **OCR** is an important means of getting text into machine-readable form, which is essential if the translator wants to make use of it, for example to develop a translation memory, or to use as a resource for searching for terminology.

Chapter 7
Corpora and the translator
Sara Laviosa

✍**1.** According to John Sinclair, a **corpus** is "… a collection of texts assumed to be representative of a given language, dialect or other subset of a language, to be used for linguistic analysis." (Sinclair, 1992: 2)

✍**2.** A **corpus** is "… a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language. (EAGLES 1996)

✍**3. Types of translation corpus**:



✍**4.** A **bilingual mono-directional parallel** corpus consists of one or more texts in language A and its/their translation(s) in language B, while a **bi-directional parallel** corpus consists also of one or more texts in language B and its/their translation(s) in language A.

✍**5.** A **bilingual comparable corpus** consists of two collections of original texts in language A and language B. The two collections are generally similar with regard to text genre, topic, time span, and communicative function.

✍ **6.** A **monolingual comparable corpus** consists of two collections of texts in one language. One collection is made up of translations from one source language (mono-SL) or a variety of source languages (multi-SL), the other consists of original texts of similar composition to the translational component.

✍ **7.** There are at least two ways in which the practising translator can benefit from the new developments in **corpus-based** research.

- ❁ They can draw on the insights provided by **descriptive studies** into the differences and similarities between languages, the strategies adopted by translators, the patterning of translational language independently of the influence of the source language, as well as the most common translation equivalents. These insights can not only enhance translation performance in terms of fluency and accuracy, but will enable them to refine their awareness of the nature of translation as a particular type of language mediation.
- ❄ On the other hand, the availability of user friendly and relatively inexpensive **software** for the automatic processing of texts as well as the accessibility of corpora on the World Wide Web may encourage translators to carry out their own linguistic, stylistic and textual analyses of single input texts or corpora for their individual needs. This will empower the translator, who will be in a position to integrate the skills and knowledge of the researcher and the practitioner and so be able to bridge the timely gap between scholarly and professional work.

Chapter 8

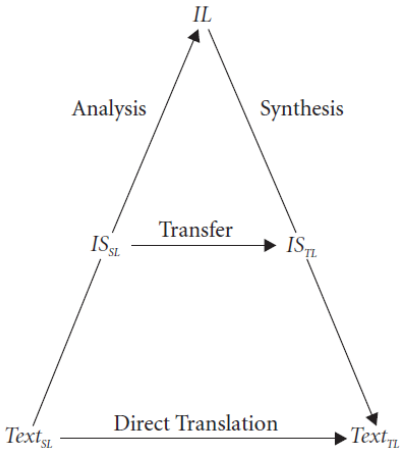
Why translation is difficult for computers

Doug Arnold

✍ **1.** Four particular **limitations of computers**; the inability of computers to:

- ✰ perform vaguely specified tasks
- ✠ learn things (as opposed to being told them)
- ✷ perform common-sense reasoning
- ✪ deal with some problems where there is a large number of potential solutions.

✍ **2.** There are three "classical" architectures for MT. These, and the tasks they involve, can most easily be understood in relation to a picture like the well-known "**pyramid diagram**" in the following figure, probably first used by **Vauquois** (1968):



✍ **3.** The simplest approach to translation is the so-called **direct approach**. Here the aim is to go directly from the source-language text to a target-language text essentially without assigning any linguistic structure. Since no structure is assigned, translation has to proceed on a word by word basis. Examples where this goes wrong are all too easy to find, and we will have little more to say about the approach.

✍ **4.** A more promising approach is base on the so-called **transfer architecture**. Here translation involves three main tasks:

- ◎ **Analysis**, where the source text is analysed to produce to an abstract representation or "interface structure" (IS) for the source-language text (ISSL). This typically contains some properties of the source language (e.g. the sourcelanguage words).
- ⌘ **Transfer**, where the source-language representation is mapped to a similar representation of the target-language text (ISTL).
- ✺ **Synthesis**, or generation, where the target-language representation is mapped to a target text.

✍ **5.** The third classical approach involves an **interlingual architecture**. Here the idea is that one has at one's disposal an "**interlingua**": a more or less language-independent representation scheme. The role of the analysis component is to produce an interlingual representation (IL), which the synthesis component can map to a target language text.

✍ **6.** A simple way to understand the relationship between these approaches is to start with the three tasks involved in the **transfer** approach, and say that the **interlingual** approach tries to eliminate the transfer task, and the **direct** approach tries to do without analysis and synthesis (i.e. it reduces everything to the transfer task).

7. This division into three tasks provides a rough classification of problems for what follows:

- Form under-determines content. That is, it is not always easy to work out the intended content from what is written. This is the **Analysis Problem**.
- Content under-determines form. That is, it is difficult to work out how a particular content should be expressed (because there is more than one way to say the same thing in any language). We will call this the **Synthesis Problem**.
- Languages differ. That is, that there are irreducible differences in the way the same content is expressed in different languages. We will call this the **Transfer Problem**, since in a transfer-based system it is typically where this problem shows up.
- Building a translation system involves a huge amount of knowledge, which must be gathered, described, and represented in a usable form. We will call this the **Problem of Description**.

8. The task of an **analysis component** is to take a source-language text (e.g. a sentence), and produce an abstract **representation**—the idea being that it will be easier to translate from this representation than from an unstructured string of source-language words. There will be different views on what sort of representation this should be (e.g. how abstract it should be), but it clearly must represent the "**content**" of the source text, since this is what the source text and its translation have in common. The problem is to infer the content from the source text. There are two major difficulties:

- The source text will often contain sentences that are ill-formed, at least from the view point of the rules in an analysis component. Analysis components must be able to cope with this by being robust.
- The source text will often be ambiguous, so it may be difficult to work out what content is intended: the form of the input under-determines its content.

9. The task of a **transfer component** is to take the sort of abstract representation produced by the source-language analysis component (call this a "source IS"), and produce something that can be input to the synthesis component of the target language (call this a "target IS"). Obviously, the closer the two ISs, the easier this will be. The transfer problem is that they cannot be the same, because languages do not associate form and content in the same ways. Thus, rules must be written to relate source and target ISs.

10. The two aspects of the **synthesis problem** are actually instances of the last problem discussed in the previous section. There are typically many

ways in which the same content can be expressed. In short: meaning under-determines form.

- The first aspect of the problem is that sometimes only one of the ways of expressing the content is correct.
- The second aspect of the synthesis problem is in some ways the converse of the first. It occurs when there is no obvious way of selecting the right way to express the content.

11. To take a very simple **example**, the content of the sentence {*Sam saw a black cat*} can be expressed in English in many other ways {**a.** Sam saw a cat. It was black. **b.** Sam saw something black. It was a cat. **c.** Sam saw a cat which was black. **d.** Sam saw a black thing which was a cat. **e.** A black cat was seen by Sam. **f.** Something happened in the past. Sam saw a cat. **g.** There was a black cat. Sam saw it.} The **problem** is how to **select** among these alternatives. In part, this is just another combinatorial problem: there are just **too many alternatives** to consider. But more serious is the problem that it is hard to know in general when one way of saying something is better than another. The only ***reliable test*** is to read what has been produced, and see if it is clear, and would be clear to a potential reader. But this is certainly asking too much of a computer. We would be asking not only that it understand sentences, but also that it should be able to consider whether someone else would be able to understand them.

12. One approach to this problem is to say "choose the output that is most similar to the source text." This is, in fact, one of the ideas behind a transfer-based approach using fairly superficial structures: by staying close to the surface, surface information from the source language is preserved, and the **synthesis problem** is made easier. But this will also lead to there being more differences between source and target language structure (cf. the transfer problem).

13. The **analysis**–**transfer**–**synthesis** approach requires an analysis and synthesis component for each language, and a transfer component for each pair of languages. For **n** languages, there are **n×(n–1)** such pairs (not $\mathbf{n^2}$, because we do not need a transfer component from any language into itself). Of course, one may expect that a lot of the transfer rules taking English to French may be workable in reverse. So one may be able to divide this number by 2. Nine languages still need 36 transfer components, 20 languages need 190 transfer components.

14. The **lexicon** contains a description of all the basic words the system is to deal with (their grammatical category, spelling, what they correspond to in the abstract representation), what complements they take (e.g.

whether they are transitive or intransitive), any idiosyncrasies of syntax or morphology.

✍**15.** The **morphological rules** describe the ways in which different forms of words are formed (e.g. plural formation: boy → boys, child → children) and the ways in which new words can be formed, e.g. by compounding (combining two independent words like film and society to make film society) or affixation (adding -ize to legal to make legalize, and then adding -ation to make legalization).

✍**16.** The **syntactic/semantic rules** describe the way in which words and phrases can be combined together to make larger phrases. Of course, in each case, the rules have to specify not only what can be combined with what, but what sort of abstract representation should be built.

✍**17.** In a reasonably sized system, one will certainly be dealing with tens of thousands of words, and with several hundred **morphological** and **syntactic** rules. Even leaving aside the fact that writing some of these rules requires fundamental research, one is clearly looking at tens of person years of effort by highly trained linguists for each language just to describe the requisite linguistic knowledge. There are three **ways** of trying to **minimize this problem**:

❶ Restrict the coverage of MT systems to very specialized domains, where vocabulary is small and the grammar is relatively simple.

❷ Exploit existing sources of knowledge, for example automatically converting machine-readable versions of monolingual or bilingual dictionaries for use in MT systems.

❸ Try to manage without explicit representations of linguistic (or non-linguistic) knowledge at all.

✍**18. The first solution** is attractive in theory, and has proved successful in practice (cf. the outstanding success of *Météo*), but its value is limited by the number of such domains that exist (it has proved very difficult to think of other domains that are as tractable as weather reports). The problem with **the second solution** is that existing dictionaries and grammars have normally been created with human users in mind, and so do not contain the kind or level of information required for use in MT. **The third solution** underlies one of the recent approaches which are discussed in the following section.

✍**19.** The last decade has seen the emergence of so-called **analogical approaches** to MT, which, at least in their radical form, dispense with the representations and rules. The analogical approaches in question are **example-based** approaches and **stochastic** or **statistical** approaches.

✍**20.** The leading idea behind so-called **Example-based MT** (EBMT) approaches is that instead of being based on rules, translation should be based on a **database of examples**, that is, pairings of fragments of source- and target-language text.

✍**21.** The intuitive appeal of **statistical** approaches can be seen when one considers how one normally approaches very complex processes involving a large number of interacting factors. One approach is to try to disentangle the various factors, describe them individually, and model their interaction.

✍**22.** Of course, there are many ways one could try to apply statistical methods in a "**classical**" approach to MT, but a more radical idea has also been proposed. The central idea is this. When presented with a French sentence f, we imagine that the original writer actually had in mind an English sentence e, but that e was somehow garbled in translation so that it came out as f. The job of the MT system is just to produce e when presented with f. Seen in this way, translation is an instance of transmission down a **noisy channel** (like a telephone line), and there is a standard technique that can be used to recover the original input (the English sentence e), at least most of the time. The idea is that f is more or less likely to occur depending on which English sentence the writer had in mind. Clearly, we want the one(s) that give f the highest **probability**.

Chapter 9

The relevance of linguistics for machine translation

Paul Bennett

✍**1.** Linguistics is concerned with providing descriptions of languages, theories of human language in general, and **formalisms** within which these descriptions and theories can be stated.

✍**2.** At least two main "**schools**" can be distinguished in linguistics.

- ✯ In **formal approaches**, the emphasis is on explicit description of the structure and meaning of words and sentences. Noam Chomsky's theory of "**generative grammar**" is the best known representative of this school.
- ✠ In contrast, **functional approaches** are more concerned with the use of language and the ways in which sentences are combined together to produce a well-formed text. Formal frameworks are far more easily incorporated into software (more computationally tractable) than functional ones, and have been more influential in MT research and development.

✍**3.** Within the **grammar**, three different levels of description can be distinguished. **Morphology** is concerned with word structure, **syntax** with sentence structure, and **semantics** with meaning.

✍**4.** It is **syntax** and **semantics** that form the core of MT systems and will therefore be the focus of attention in what follows, but morphology also raises a number of translation problems, as seen, for instance, in novel words like transferee, *Murdochization* and *dome fiasco*.

✍**5.** There are also two fundamental ways in which linguistics can feed into MT research.

- ✷ The first is by way of **representations**. Whether in a transfer-based or interlingua system, a sentence is converted to some representation of its structure or meaning. Linguistic notions can play a crucial role in determining what such representations look like and what representation is appropriate for a particular example.
- ✪ The other is in terms of **description**, i.e. the modules of the system which describe or capture various kinds of knowledge—the grammars, lexicons or transfer components.

Chapter 10

Commercial systems: The state of the art

John Hutchins

✍**1.** There are numerous examples of the successful and long-term use of MT systems by multinationals for technical documentation. One of the best known has been the application of the **Logos** system at the Lexi-Tech company in New Brunswick, Canada; initially for the translation into French of manuals for the maintenance of naval frigates, later as a service for many other large translation projects. **Systran** has had many large clients: Ford, General Motors, Aérospatiale, Berlitz, Xerox, etc.

✍**2.** In the 1990s, the options for large-scale computer-based translation production broadened with the appearance on the market of translator's **workstations**. These combine multilingual word processing, means of receiving and sending electronic documents, facilities for document scanning by **OCR**, terminology management software, facilities for concordancing, and in particular **TMs**. The latter facility enables translators to store original texts and their translated versions side by side, so that corresponding sentences of the source and target are **aligned**. The translator can thus search for a phrase or even full sentence in one language in the **TM** and have displayed corresponding phrases in the other language. These may be either exact matches or approximations ranked according to closeness.

3. One of the fastest growing areas for the use of computers in translation is **software localisation**. Here the demand is for producing documentation in many languages to be available at the time of the launch of new software. Translation has to be done quickly, but there is much repetition of information from one version to another. **MT** and, more recently, **TMs** in translator's workstations are the obvious solution. Among the first in this field was the large software company *SAP AG* in Germany, using older **MT** systems, Metal and Logos. Most **localisation**, however, is based on the **TM** and workstation approach—mainly Transit, Déjà Vu, and the Trados Workbench.

4. Systems for **assimilation purposes** (for the less-demanding "occasional" user) are also widely available, with good language coverage on the whole. However, these systems often give **poor-quality** output, even for well-written source texts, let alone the low-level writing on e-mail and other Internet applications.

Chapter 11

Inside commercial machine translation

Scott Bennett and Laurie Gerber

1. Methods for building MT systems may be classified by their position on a continuum between two extremes:

- Manually created systems where the lexicon, grammar and translation rules are written by linguists. We will call these "**rule-based**" systems.
- Systems where patterns are learned automatically by the computer from texts. We will call these "**data-driven**" systems.

2. **Rule-based** MT developers have internally defined proprietary grammars, and **symbolic representations**. The grammar allows linguists to catalog the types of linguistic phenomena that the system needs to use. When planning an MT system for a new language pair, the job of linguists and engineers is to identify appropriate mappings and parsing techniques between the set of phenomena realized in the new source and/or target languages and the system's grammar. The symbolic representation is the data structure in the computer that holds all of the grammatical information about a unit of text, and allows the parser to add incrementally new information as it is discovered, and query the information already stored. The **unit** of translation is usually a **sentence**.

3. One of the first challenges encountered when developing a **rule-based** MT system is where to find the resources—grammatical information about the languages involved, example texts for translation, lists of words

and terms, and reliable translation equivalents for words and phrases. General-purpose systems, such as ***Logos*** and ***Systran***, may be used on any type of text from any domain. This means that these systems must come equipped with a large general vocabulary, and that development work for production use must be grounded in observation and testing of extensive real-world text.

4. The "**translation rules**" learned by statistical systems consist of "**parameters**", cross-lingual correspondences between words or phrases, accompanied by the probability that the word or phrase in the source language will be rendered as the word or phrase in the target language. In order to build such a system, sentence-to-sentence correspondences must be established, and words separated from punctuation, or "**tokenized**". It is this aligned, tokenized "parallel corpus" that a statistical system learns from.

5. All MT developers have, as do ***Logos*** and ***Systran***, internally determined performance thresholds for product release. Preparation for release includes an objective **evaluation** of system output: either on a targeted task (if the system is developed for a particular domain or text type), or on a balanced corpus representing various text types (if the MT system is intended to be a "*general purpose*" system).

6. The **type of use** a system is *targeted for* includes whether it will be primarily applied to

- **assimilation** or gisting tasks (information gathering, and browsing, where speed, and broad lexical coverage are more important than quality),
- **dissemination** tasks (translation for publication, where quality is most important, but the user has authoring control and may employ controlled language or at least work with a limited vocabulary and text type), or
- **communication** tasks (real-time e-mail translation, for example, where speed and accuracy are both important, as is the ability to handle informal language, but where extensive technical terms are unlikely to appear).

Chapter 12

Going live on the internet

Jin Yang and Elke Lange

1. The **AltaVista** translation service with ***Systran*** is a good showcase for MT technology. The explosive and positive user feedback shows that MT has proven its worth in practice. Improving **translation quality** and expanding **language coverage** are definitely pressing challenges.

Chapter 13
How to evaluate machine translation
John S. White

1. Fidelity and **intelligibility** are, of course, correlated: a completely unintelligible expression conveys no information.

2. Evaluation types:
- ◈ Feasibility evaluation
- ✵ Internal evaluation
- ❁ Declarative evaluation
- ⌖ Usability evaluation
- ❄ Operational evaluation
- ▣ Comparison evaluation

3. The very first glimpse the general public got of MT was essentially the result of a **feasibility** study, that is, an evaluation of the possibility for a particular feat to be accomplished at all, or for a particular approach, whether it has any actual potential for success after further research and implementation. **Feasibility evaluations** provide measures of interest to researchers and the sponsors of research.

4. Internal evaluation occurs on a continual or periodic basis in the course of research and or development. Here, the question is whether the components of an experimental, prototype, or pre-release system work as they are intended. The particular items covered in such an evaluation will vary with the maturity of the system being evaluated of course, and thus provide measures of interest to researchers, research sponsors, developers, and vendors. As with the feasibility test, we want to be able to show that we can cover the fundamental contrastive phenomena of the ***language pair***. But we need to show some other attributes as well, namely that the system we are developing, or bringing to market, or adapting to our own user environment, is improving. We need to show, for instance, that as we add ***grammar rules***, or dictionary entries, the system translates the things we are trying to improve better than it did, and does not suddenly fail to do something it used to do. So we need to have a standard set of test materials for **iterative testing** (tests designed to make sure an improvement in one area actually works and does not adversely affect another area).

5. The way you look at the relationship of the **input** and **output** has been referred to as the difference between "**black-box**" testing and "**glass-box**" testing.

- ☆ The **black-box** view is a look at the input and output without taking into account the mechanics of the translation engine.
- ✠ The **glass-box** view looks inside the translation engine to see if its components each did what was expected of them in the course of the translation process.

✍**6.** There are **advantages** to each:

- ✷ The **black-box** view is portable (i.e., the method and measures are external to the design and philosophy of any one system). It is more amenable to comparisons of systems, and to determining the current language coverage of a particular system.
- ✪ The **glass-box** view helps to determine the extensibility of coverage of the system, by being able to tell whether and how well the designed processes perform their functions.

✍**7. Declarative evaluation** is the heart of the matter for the casual observer. It addresses the question of whether a system translates well, by which is meant, among other things, the degree to which it has the attributes of **fidelity** and **intelligibility** that we introduced above. This evaluation type is clearly of particular value to investors, end-users, vendors, and managers, but also to developers. The purpose of **declarative evaluation** is to measure the ability of an MT system to handle text representative of actual end use.

✍**8. Declarative evaluation**: ***ALPAC*** (1966): This evaluation designed by John B. Carroll comes from the early days of MT, and is described in the **ALPAC report** we have already introduced. Carroll sought a standard method of evaluating both human and machine translation, that was simple and portable, yet highly reliable. He realized that subjective judgments about translations show promise of meeting these goals. He also realized all of the human factors that come with ***subjective*** judgments. The method he arrived at is an ingenious optimisation of simplicity and portability, while incorporating as many controls against human biases as were possible and practical.

✍**9.** The purpose of a **usability evaluation** is to measure the ability of a system to be useful to people whose expertise lies outside MT *per se*. As we have described the user set above, these people may be translators, editors, analysts requiring a particular type of information, or any other sort of information consumer.

✍**10.** The **usability** of a system is a function of two attributes, the **utility** of an application and the users' **satisfaction** with it.

✍**11. Usability** is measured at the point of interaction between the user and the thing being used, in this case the MT software application, and this

means that the focus of such evaluation is on the apparent functioning of the user interface. **Evaluation of interface** properties may include:

- the time to complete a particular task
- the number of steps to complete it
- how natural the navigation process appears to be
- how easy it is to learn how to use the application
- how helpful the documentation is

12. Operational evaluations answer the question "Is it worth it?". Here, the primary factors to consider are all of the costs involved, against all of the benefits. Issues like common platforms and operating systems are germane here. End users and their managers need these evaluations, and thus investors and vendors must be attentive to the operational factors. The ***purpose*** of **operational evaluation** is to determine the cost-effectiveness of an MT system in the context of a particular operational environment.

13. A meaningful measure in **operational** evaluation is **return on investment**, which implies comparison of the measurement of the real costs of an MT application, and the real benefit (revenue, cost savings, etc.). We then may compare the value of these properties against the same measurements of the way the process is currently done. The result is an expression of the benefits of inserting MT technology (or not), expressed in terms of the attributes of productivity, cost, revenue, or quality.

14. Among the factors to be considered in measuring these attributes are these:

- **Operational environment**
  - – compatibility with the familiar (Does the MT software (appear to) run on my desktop computer?);
  - – compatibility with the standard formats (Does the MT system accept input from, and output to, the OA formats I use everyday?);
  - – consistency of the application GUI with the operating system (Are the common toolbar items in the same place in this application as they are in the other applications I use?);
  - – response time (less an operational issue than it once was, and perhaps more of a usability issue: Does it have roughly the same response time as the other applications I use?);
  - – humans in the loop (Does this application require human intervention to prepare/correct data, or to operate the application?);
  - – preparation, throughput, correction, and output times.
- **Application Design**
  - – extensibility (Does the system have a user-accessible lexicon, or other ways to customize for this environment?);

– use of standards (e.g., Does it handle the common codes for writing systems?);
– number of steps to complete a task (i.e., the number of steps designed or recommended);
– fail-softs (Does an MT failure cause an exit from the program? Does it cause a system crash?).

❁ **Provider**

– documentation (Is it complete and helpful?);
– support (Is the support timely and adequate?);
– improvement (Are there periodic new releases? Do they fix user-discovered bugs?);
– corporate situation of provider (Will the provider be around long enough to support the product system trough its life cycle?).

⌘ **Cost**

– of the system (hardware, software, licenses);
– of maintenance;
– of the process (both the automatic parts and the human intervention parts);
– of human translation (i.e., Does the overall MT process wind up being cheaper than professional human translation?).

✍ **15. Comparisons measure** some attribute of a system against the same attributes of other systems. Thus the methods of comparison are the same as the methods of the other evaluation types, applied among several systems. This is of obvious benefit to purchasers of systems and investors in system development and productization.

✍ **16.** The *purpose* of **comparison evaluations** is to determine the best system, best implementation, or even the best theoretical approach for meeting current or future needs. It appears that comparison evaluation can measure the same attributes as the feasibility, internal, operational—in fact, any of the other types. Depending on what we are comparing, it has all of the properties of any of these other types, except that in each case we are holding the measurements of one against the same measurements of another.

✍ **17.** During the 1990s, the US government Defense Advanced Research Projects Agency (**DARPA**) developed a set of methods for evaluating MT which sought to express meaningful measures of the performance of the system prototypes of its three funded MT projects. There was a **big problem**, though, namely, the **three projects** had very **little in common**.

✠ Each system translated different language pairs (French, Spanish, and Japanese into English).
❄ Each system envisioned a different end-use environment (automatic batch translation vs. human-interactive translation vs. authoring tools).

- Each project had radically different theoretical approaches to translation, from purely statistic to purely knowledge driven, and points in between.

18. The **DARPA** methods could **not** take advantage of any linguistic phenomena (because of the different pairs involved), or anything in common about the system's approaches (since the approaches are so different). This was the ultimate "**black-box**" requirement.

19. The **DARPA** methods used the judgments of target native speakers, who did not know the source languages, to make a variety of judgments about intelligibility and fidelity through three exercises:

- **Adequacy:** this is a fidelity measure intended to capture how much of the original content of a text is conveyed, regardless of how imperfect the English output might be. In this evaluation, expert human translations were divided up into syntactic "chunks", and then arranged side by side with a system translation (without any chunks). The English speakers ("evaluators") were asked to look at each fragment, and indicate on a 1–5 scale the degree to which the information in the fragment is present in the translation.
- **Fluency.** This is an intelligibility measure, designed to determine how much like "good English" a translation appears to be, without knowing anything about what information is supposed to be there. Here, evaluators used another 1–5 scale to judge documents a sentence at a time.
- **Informativeness.** This is another fidelity measure, used to determine whether there is enough information in the translation to answer specific questions about its content. Evaluators answer multiple-choice questions about a translation rather like a reading comprehension test (except that we are testing the reading and not the reader).

20. One of the things we might say in common about all of the **evaluation types** is that their methods must be designed and done carefully, to control for the sources of variance. Most of the types take time, effort, and coordination to perform. Some way to automate some or all of the evaluation types would be extremely beneficial for the field, allowing for the critical choices of all the stakeholders to be made much more rapidly and consistently. For some types, e.g., usability, automated measurement may be possible today. For **declarative** and **internal** evaluations, automation is much harder because of the "**ground truth**" problem that translation has. A solution may lie in discovering consistent correlations between the attributes we need to measure and measurements we can make automatically.

Chapter 14

Controlled language for authoring and translation

Eric Nyberg, Teruko Mitamura & Willem-Olaf Huijsen

✍ **1.** A **CL** is an explicitly defined restriction of a natural language that specifies constraints on lexicon, grammar, and style.

✍ **2.** It is important to note that there is **no single CL**, say for English, which is approved by some global authority. In practice, there are several different definitions of CL, which are proposed by individual groups of users or organizations for different types of documents.

✍ **3. CL** can be used solely as a guideline for authoring, with self-imposed conformance on the part of the writer; **CL** can be used with software which performs a complete check of each new text to verify conformance; and **CL** can also be incorporated into a system for automatic MT of technical text. In all cases, the overall aim is to reduce the ambiguity and complexity of the text, whether it is processed by machine or read by humans only.

✍ **4. CLs** can be characterized as human-oriented or machine-oriented.

- ✪ **Human-oriented** CLs intend to improve text comprehension by humans;
- ❋ **Machine-oriented** CLs intend to improve "text comprehension" by computers.

✍ **5.** The general **advantage** of CLs is that they make many aspects of text manipulation easier for both humans and computer programs. The reduction in homonymy, synonymy, and complexity of the lexicon and the adherence to writing rules may improve the **readability** and **comprehensibility** of the text. Consequently, the performance of tasks that involve the documentation can be more efficient and effective. This **advantage** is especially relevant for complex texts, and also for nonnative speakers. All documents written in the **CL** will exhibit a ***uniformity in word choice***, use of terminology, sentence structure, and style, which makes them easier to maintain and reuse. It is also the case that the use of CL improves both the **consistency** and **reusability** of the source text.

✍ **6.** The use of **CLs** also has a number of **potential drawbacks** however. From the author's point of view, the writing task may become more time-consuming. It can take more concentration to write documents if they must conform to the rules of a **CL**, which can slow down the writing process. **CLs** which are not supported by automatic checking require self-vigilance on the part of the author, which can also be time-consuming. Rewriting a sentence which does not conform is often more

complex than the simple substitution of approved counterparts for unapproved words, and sometimes requires rewriting the whole sentence.

7. CL checkers are programs which assist authors in determining whether their text complies with the specification of a CL. This assistance is generally given as a series of **critiques** or issues that are raised with respect to the text, communicated to the user as text messages by the software.

8. MT is potentially one of the most interesting computational applications of **CL**. If a CL and an **MT** system are attuned to each other, **MT** of texts written in that **CL** can be much more efficient and effective, requiring far less—or ideally even no—human intervention.

9. CL for **MT** works well when the following characteristics are present in the intended application domain:

- **Translation for dissemination**. When documents are authored in one language, in a particular domain, and are then translated into multiple languages, it is possible to control the style and content of the source text. This type of translation is referred to as "translation for dissemination". A given domain is less amenable to a CL approach when unrestricted texts from multiple source languages are to be translated into one target language. This type of translation is referred to as "translation for assimilation".
- **Highly-trained authors**. It may not be easy to deploy CL in an existing authoring process at first, because authors are used to writing texts in their own style for many years. Therefore, it is crucial for success that the authors are able to accept the notion of CL, and are willing to receive CL training. It seems that authors who receive comprehensive training and who use CL on a daily basis achieve the best results and highest productivity. It is also important that these well-trained authors act as mentors during the training of other authors new to CL. Adequate training and mentoring is crucial for author acceptance of CL.
- **Use of CL checkers**. Although CL can be implemented simply as a set of written guidelines for authors, uniform quality of CL text is maximized if the author uses a CL checker to write texts which are verified to comply with the CL definition. The use of an on-line checking system enhances consistency and promotes the reuse of texts across similar product lines where appropriate. Authored texts can also be aligned with their translations in a translation memory, leading to increases in production efficiency for technical authoring and translation.
- **Well-defined domain**. The success of a CL relies heavily on ruling out ambiguous meanings for terms which are not required in the given

domain. Therefore, CL may be less suitable for unrestricted domains, such as general newsletters, email or bulletins. On the other hand, it is possible to control technical vocabulary and writing style in most technical documentation, since the domain is specific and it is preferable to standardize terminology and writing style.

✍ **10.** A **CL** for **MT** attempts to rule out difficult sentence structures and to limit ambiguous vocabulary items in order to achieve accurate translation. However, if a **CL** becomes too restrictive, it may introduce usability and productivity problems. If it is too difficult to write sentences that comply with the **CL**, no one will use it. **Controlled sentences** which are not stylistically adequate will not be accepted by authors and will be heavily post-edited by translators. Therefore, it is essential to find a middle ground which is productive and acceptable for authors and which promotes high-quality translation. In order to improve author productivity, it is desirable to develop an automatic rewriting system to convert text into **CL**. For the field of **CL**, this will be a new challenge and a future direction of research and development.

Chapter 15
Sublanguage
Harold Somers

✍ **1.** The term **sublanguage**, usually used in connection with MT, dates back to Zellig Harris, the structuralist linguist, who gave a precise characterization of the idea in terms of his linguistic theory. The term was coined with the mathematical idea of "**subsystem**" in mind, the "**sub-**" prefix indicating not inferiority, but inclusion. So a **sublanguage** is a subset of the "whole" language.

✍ **2.** Like **controlled language**, a **sublanguage approach** to MT (and many other computational linguistics tasks) benefits from the two main characteristics of sublanguage as compared to the whole language, namely the reduced requirement of coverage in the lexicon and grammar.

✍ **3.** The term **sublanguage** has come to be used … for those sets of sentences whose lexical and grammatical restrictions reflect the restricted sets of objects and relations found in a given domain of discourse. (Kittredge and Lehrberger, 1982: 2)

✍ **4.** A **sublanguage** arises when a community of users—domain specialists—communicate amongst themselves. They develop their own vocabulary, that is not only specialist terms which have no meaning to outsiders, but also (and crucially) everyday words are given narrower interpretations, corresponding to the concepts that characterize and

define the domain. In addition, there will be a favoured "**style**" of writing or speaking, with preferred grammatical usages.

Chapter 16
Post-editing
Jeffrey Allen

1. Post-editing is by far most commonly associated as a task related to MT and has been previously defined as the "term used for the correction of machine translation output by human linguists/editors" (Veale and Way, 1997).

2. Another good summary statement indicates that "**post-editing** entails correction of a pre-translated text rather than translation 'from scratch'" (Wagner, 1985).

3. In basic terms, the task of the **post-editor** is to edit, modify and/or correct pre-translated text that has been processed by an MT system from a source language into (a) target language(s).

4. Pre-editing and controlled language writing principles are often used in tandem with the post-editing approach in order to improve the translatability of technical texts and to speed up the productivity of the **post-editing** process.

5. The level of **post-editing** to be performed on a text is entirely dependent on several factors, including:

- the user/client,
- the volume of documentation expected to be processed,
- the expectation with regard to the level of quality for reading the final draft of the translated product,
- the translation turn-around time,
- the use of the document with regard to the life expectancy and perishability of the information,
- the use of the final text in the range from information gisting to publishable information.

6. Minimal post-editing is a fuzzy, wide-range category because it often depends on how the post-editors define and implement the "minimum" amount of changes to make in view of the client/reader audience.

Chapter 17
Machine translation in the classroom
Harold Somers

1. The most interesting aspect of **MT** for **CL** is that, more than any other application, translation requires "**coverage**" of all the linguistic levels in more than one language. For this reason **MT** is sometimes seen as the archetypical application of **CL**. Another useful feature of translation as a test-bed for **CL** techniques is that you can usually tell pretty well whether an MT program has "**worked**" (notwithstanding subtle difficulties of saying just how "**good**" a translation is, it is usually quite clear whether some piece of text is or is not a translation of another text).

2. For the **student** (and teacher) of **CL**, then, **translation software** can be used to illustrate problems (and solutions) in **language analysis** at various levels both monolingually and contrastively. Source-text analysis requires morphological disambiguation (*is a tower a high structure or something that tows?*) and interpretation (*is books the plural of book, or a form of the verb to book?*), word-sense disambiguation (*bank: financial institution or side of a river?*), syntactic, semantic and pragmatic disambiguation. Translation involves **converting** linguistic aspects of the source text into their appropriate form in the target text, thus the application of contrastive lexical and syntactic knowledge. And the **generation of the target text** involves the corresponding problems of style, syntax, and morphology.

3. More generally, **translation software** output can be used with students of CL for linguistic error analysis in general or focussing on one particular problem area, using a specially designed **test suite**. For example, if one was interested in the subtleties of modality (in English, expressed by words like can, must, should, ought to, etc.) one could construct a set of sentences which express different modalities, and see how they are translated. Other interesting linguistic phenomena which illuminate contrastive differences between languages are the use of tenses, (in)definiteness, passive constructions and other means of **topicalisation**, and so on.

Short Answer Items & Tests

4.2 Short Answer Items

1. The example database of a translation memory is an example of a corpus, with the particularly interesting property of being an ……. corpus, by which is meant that it represents texts which are translations of each other, and, crucially, the corpus has been subdivided into smaller fragments which correspond to each other.

2. ……. matching allows translators to retrieve records for morphological variants, for spelling variants, and for multiword terms, even if the translators do not know the precise order of the elements in the multiword term.

3. Term-extraction tools can be either ……. or ……. .

4. Term-extraction tools that use a ……. approach basically look for repeated sequences of lexical items.

5. The ……. threshold, which refers to the number of times that a sequence of words must be repeated, can often be specified by the user.

4.3 Answers

1) aligned parallel	**2)** Fuzzy
3) monolingual, bilingual	**4)** statistical
5) frequency	

☙☆❧ 4.4 Tests ☙☆❧

✍ Select the best choice.

1. A prerequisite for a TM system is of course a database of translation examples. Known to computational linguists as an "……. corpus", there are principally three ways of building a TM database: building it up as you go along, importing it from elsewhere, or creating it from a parallel text.

a) non-aligned parallel
b) non-aligned serial
c) aligned serial
d) aligned parallel

2. In ……., ……., there is an aligned parallel corpus of previous translations, and from this corpus are selected appropriate matches to the given input sentence. In a(n) ……., however, it is up to the user, the translator, to decide what to do with the retrieved matches.

a) SMT, like in TMs, EBMT
b) EBMT, in contrast to TMs, TM
c) EBMT, like in TMs, TM
d) SMT, in contrast to TMs, EBMT

3. In ……., we try to automate the process of selecting the best matches or fragments from the best matches, and then to "recombine" the corresponding target-language fragments to form the translation.

a) EBMT
b) EBMT and TM
c) TM and SMT (but not in EBMT)
d) TM (but not in SMT and EBMT)

4. Some sophisticated TMSs employ ……. matching techniques. A ……. match will retrieve those term records that are similar to the requested search pattern, but which do not match it exactly.

a) fuzzy, fuzzy
b) fuzzy, lucid
c) lucid, wildcard
d) wildcard, cogent

5. In cases where ……. searching or ……. matching is used, it is possible that more than one record will be retrieved as a potential match. When this happens, translators are presented with a ……. list of all the records in the ……. that may be of interest and they can select the record(s) they wish to view.

a) termbase, fuzzy, hit, wildcard
b) wildcard, hit, fuzzy, termbase
c) hit, fuzzy, termbase, wildcard
d) wildcard, fuzzy, hit, termbase

☙✷❧ 4.5 Answer key ☙✷❧

	a	b	c	d		a	b	c	d
1				×	2			×	
3	×				4	×			
5				×					

Book ❺

Learning Machine Translation

Cyril Goutte *Nicola Cancedda*
Marc Dymetman *George Foster*

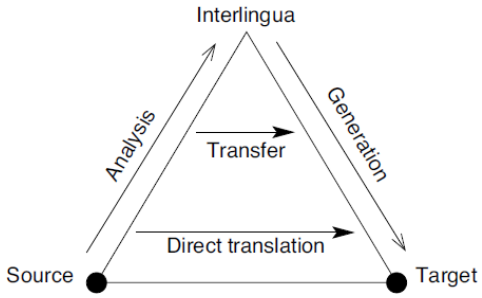
5.1 Notes

Chapter 1

A Statistical Machine Translation Primer

Nicola Cancedda, Marc Dymetman, George Foster, and Cyril Goutte

✍ **1.** The machine translation pyramid. Approaches vary depending on how much analysis and generation is needed. The **interlingua** approach does full analysis and generation, whereas the **direct** translation approach does a minimum of analysis and generation. The **transfer** approach is somewhere in between.



✍ **2.** The general setting of **statistical machine translation** is to learn how to translate from a large corpus of pairs of equivalent source and target sentences. This is typically a machine learning framework: we have an **input** (the source sentence), an **output** (the target sentence), and a **model** trying to produce the correct output for each given input.

✍ **3.** The early approach to **SMT** advocated by the IBM group relies on the **source-channel approach**. This is essentially a framework for combining two models: a word-based **translation model** and a **language model**.

- The **translation model** ensures that the system produces target hypotheses that correspond to the source sentence.

- The **language model** ensures that the output is as grammatical and fluent as possible.

4. Evaluation of Machine Translation: *Levenshtein-Based Measures*: A first group of measures is inherited from **speech recognition** and is based on computing the **edit distance** between the candidate translation and the reference. This distance can be computed using simple dynamic programming algorithms.

5. ***Word error rate (WER)*** (Niesen et al., 2000) is the sum of insertions, deletions, and substitutions normalized by the length of the reference sentence. A slight variant (WERg) normalizes this value by the length of the Levenshtein path, i.e., the sum of insertions, deletions, substitutions, and matches: this ensures that the measure is between zero (when the produced sentence is identical to the reference) and one (when the candidate must be entirely deleted, and all words in the reference must be inserted).

6. ***Position-independent word error rate (PER)*** (Tillmann et al., 1997b) is a variant that does not take into account the relative position of words: it simply computes the size of the intersection of the bags of words of the candidate and the reference, seen as multi-sets, and normalizes it by the size of the bag of words of the reference.

7. A second group of measures, by far the most widespread, is based on notions derived from information retrieval, applied to the n-grams of different length that appear in the candidate translation. In particular, the basic element is the **clipped n-gram precision**, i.e., the fraction of **n-grams** in a set of translated sentences that can be found in the respective references.

8. ***BLEU*** (Papineni et al., 2002) is the geometric mean of **clipped n-gram precisions** for different *n-gram lengths* (usually from one to four), multiplied by a factor (brevity penalty) that penalizes producing short sentences containing only highly reliable portions of the translation.

9. Precision is **clipped** because counts are thresholded to the number of occurrences of n-grams in the reference, so that each n-gram occurrence in the reference can be used to "**match**" at most one n-gram occurrence in the proposed sentence. Note also that the precision is computed for all n-grams in a document at once, not sentence by sentence.

10. BLEU was the starting point for a measure that was used in evaluations organized by the U.S. National Institute for Standards and Technology (**NIST**), and is thereafter referred to as the **NIST** score (Doddington, 2002). **NIST** is the arithmetic mean of clipped n-gram precisions for different n-gram lengths, also multiplied by a (different) brevity penalty.

Also, when computing the **NIST** score, n-grams are weighted according to their frequency, so that less frequent (and thus more informative) n-grams are given more weight.

✍ **11. BLEU** and **NIST** are forced to include a **brevity penalty** because they are based only on n-gram precision. **N-gram recall** was not introduced because it was not immediately obvious how to meaningfully define it in cases where multiple reference translations are available. A way to do so was presented in Melamed et al. (2003): the general text matcher (**GTM**) measure relies on first finding a maximum matching between a candidate translation and a set of references, and then computing the ratio between the size of this matching (modified to favor long matching contiguous n-grams) and the length of the translation (for precision) or the mean length of the reference (for **recall**). The harmonic mean of precision and recall can furthermore be taken to provide the **F-measure**, familiar in natural language processing.

✍ **12.** A further measure, which can be seen as a generalization of both **BLEU** and **ROUGE** (both -L and -S), is BLANC (Lita et al., 2005). In BLANC the score is computed as a weighted sum of all matches of all subsequences (i.e., **n-grams** possibly interrupted by gaps) between the candidate translation and the reference. Parameters of the scoring function can be tuned on corpora for which human judgments are available in order to improve correlation with adequacy, fluency, or any other measure that is deemed relevant.

✍ **13.** Finally, the proposers of **METEOR** (Banerjee and Lavie, 2005) put more weight on recall than on precision in the harmonic mean, as they observed that this improved correlation with human judgment. **METEOR** also allows matching words which are not identical, based on stemming and possibly on additional linguistic processing.

✍ **14.** Liu and Gildea (2005) propose a set of measures capable of taking **long-distance syntactic phenomena** into account. These measures require the candidates and the references to be syntactically analyzed. Inspired by BLEU and NIST, averaged precision of paths or subtrees in the **syntax trees** are then computed.

✍ **15.** In the same line, Gim´enez and M`arquez (2007b) also use linguistic processing, up to shallow **semantic analysis**, to extract additional statistics that are integrated in new measures.

✍ **16.** An interesting method to **combine** the complementary strengths of different measures, and at the same time evaluate evaluation measures and estimate the reliability of a test set, is **QARLA** (Gim´enez and Amig´o, 2006).

17. A **language model** (LM), in the basic sense of the term, is a computable probability distribution over word sequences, typically sentences, which attempts to approximate an underlying stochastic process on the basis of an observed corpus of sequences produced by that process.

18. Language models have many applications apart from statistical machine translation, among them: speech recognition (SR), spelling correction, handwriting recognition, optical character recognition, information retrieval. Historically, much of their development has been linked to speech recognition and often the methods developed in this context have been transposed to other areas; to a large extent this remains true today.

19. Phrase-based MT is currently the dominant approach in statistical MT. It incorporates five key innovations relative to the classic approach the use of log-linear models instead of a simple product of language and translation models;

- the use of multiword "phrases" instead of words as the basic unit of translation, within a simple one-to-one generative translation model;
- minimum error-rate training of log-linear models with respect to an automatic metric such as BLEU, instead of maximum likelihood training;
- a clearly defined and efficient heuristic Viterbi beam search procedure; and
- a second rescoring pass to select the best hypothesis from a small set of candidates identified during search.

20. The motivations for using syntax in SMT are related to consideration of **fluency** and **adequacy** of the translations produced:

- **Fluency** of output depends closely on the ability to handle such things as agreement, case markers, verb-controlled prepositions, order of arguments and modifiers relative to their head, and numerous other phenomena which are controlled by the syntax of the target language and can only be approximated by n-gram language models.
- **Adequacy** of output depends on the ability to disambiguate the input and to correctly reconstruct in the output the relative semantic roles of constituents in the input. Disambiguation is sometimes possible only on the basis of parsing the input, and reconstructing relative roles is often poorly approximated by models of reordering that penalize distortions between the source and the target word orders, as is common in phrase-based models; this problem becomes more and more severe when the source and target languages are typologically remote from each other (e.g., subject- verb-object languages such as

English, subject-object-verb languages such as Japanese, or languages that allow relatively "free" word order such as Czech).

✍ **21.** A rather radical departure from existing approaches to SMT is proposed by Wang et al. (2007). Using **kernels** on strings it is possible to map separately sentences of the source and of the target language into distinct vector spaces (or feature spaces). Conceptually the translation problem can thus be decomposed into:

- ✯ mapping a source language sentence into a vector in the input feature space;
- ❁ mapping this vector into a vector in the output feature space by means of an appropriate function;
- ◈ mapping a vector from the output feature space into a target language sentence.

✍ **22.** The function in the **second** step can be learned from a training set using an appropriate regression algorithm (such as ridge regression). In practice, the **first** and the **second** steps are conflated in that a kernel is used to implicitly map source sentences into the input feature space. The **third** step, the inverse image problem, can be very hard, depending on the kernel used on the target side.

PART I
Enabling Technologies
Chapter 2
Mining Patents for Parallel Corpora
Masao Utiyama and Hitoshi Isahara

✍ **1.** In this chapter, the authors show that a large amount of *parallel text* can be obtained by mining comparable patent corpora. This is because patents of the **same subject** matter are often filed in multiple countries. Such patents are called **patent families**.

✍ **2. Large-scale parallel corpora** are indispensable language resources for MT. However, there are only a few publicly available large-scale parallel corpora. The authors have developed a Japanese-English patent parallel corpus created from Japanese and U.S. patent data provided for the NTCIR-6 patent retrieval task.

✍ **3.** The authors used Utiyama and Isahara's method and extracted about 2 million **clean sentence alignments**. This is the largest Japanese-English parallel corpus to date. Its size is comparable to other large-scale parallel corpora. This corpus and its extension will be used in the NTCIR-7

patent MT task and made available to the public after the 7th NTCIR-7 workshop meeting.

Chapter 3
Automatic Construction of Multilingual Name Dictionaries
Bruno Pouliquen and Ralf Steinberger

1. There is an–often not explicitly mentioned–assumption that **names** do not need **translating**. To some extent, this is true, at least for person names in related languages such as those spoken in western European countries. The usefulness of name translation is much more obvious for languages using different writing systems, such as the Chinese, Arabic, and Cyrillic scripts.

2. Starting from the observation that "**name translation** has proven to be a challenge for machine translation providers," Hirschman et al. (2000) identified the following three types of problems related to proper names:

- Translation of proper names as if they were normal meaningful words (e.g., the name of the former German Chancellor Helmut Kohl translated as Helmut Cabbage).
- Idiomatic rather than literal translation of names; this mostly concerns organization names (e.g., Escuela de Derecho de Harvard should not be back-transliterated as Harvard School of the Right, but the original Harvard Law School should be used).
- Rendering of names in a format that is unusable by target language processing. This is mainly an issue of transliteration, as foreign characters (such as those of the Cyrillic alphabet) cannot be displayed or read in other languages.

3. In the past, dictionaries were developed for general or subject-specific vocabularies, but not for **proper names**. However, name dictionaries, which may include crosslingual, cross-script, and also monolingual name variants, are a precious resource that can help improve the output of many text analysis applications. These include **machine translation**, information retrieval, topic tracking, relation and event extraction, the automatic generation of social networks based on information found in free text, and more. While work on automatically extracting information to feed name dictionaries is still rather scarce, many scientists now work on automatically learning transliteration rules and name equivalences from bilingual name lists, especially for Arabic and Russian.

4. The authors have presented work on recognizing **new names** in multilingual news collections in 19 languages and on an automatic procedure to determine whether any new name is likely to be a variant of a known name or whether it is a name in its own right. For that purpose,

each name is normalized—using language pair-independent rules–and then compared to each of the known names in the database using a combination of two similarity measures. The language-independence of the rules is of particular importance because names found in news texts can come from any country and could be pronounced according to the pronunciation rules of any language on the globe.

Chapter 4

Named Entity Transliteration and Discovery in Multilingual Corpora

Alexandre Klementiev and Dan Roth

✍ **1. Named entity recognition** (NER) is an important part of many natural language processing tasks. Current approaches often employ machine learning techniques and require supervised data. However, many languages lack such resources.

✍ **2.** A major challenge inherent in discovering transliterated **NEs** is the fact that a single entity may be represented by multiple transliteration strings.

- ✠ One reason is language morphology. For example, in Russian, depending on the case being used, the same noun may appear with various endings.
- ✪ Another reason is the lack of transliteration standards. Again, in Russian, several possible transliterations of an English entity may be acceptable, as long as they are phonetically similar to the source.

✍ **3.** The authors have proposed a **novel algorithm** for cross-lingual multiword NE discovery in a bilingual weakly temporally aligned corpus. The authors have demonstrated that using two independent sources of information (**transliteration** and **temporal similarity**) together to guide NE extraction gives better performance than using either of them alone. The algorithm requires almost **no supervision** or **linguistic knowledge**. Indeed, the authors used a very small bootstrapping training set and made a simple assumption in order to group morphological variants of the same word into equivalence classes in Russian.

Chapter 5

Combination of Statistical Word Alignments Based on Multiple Preprocessing Schemes

Jakob Elming, Nizar Habash, and Josep M. Crego

✍ **1.** Although **phrase-based approaches** to SMT tend to be robust to word-alignment errors (Lopez and Resnik, 2006), improving **word alignment** is still meaningful for other NLP research that is more sensitive to

alignment quality, e.g., projection of information across parallel corpora (Yarowsky et al., 2001).

✍ **2.** The authors have presented an approach for using and combining multiple alignments created using different preprocessing schemes. Their results show that the remapping strategy improves **alignment correctness** by itself.

✍ **3.** The authors showed that the combination of multiple remappings improves **word alignment** measurably over a commonly used state-of-the-art baseline. The authors obtained a relative reduction of alignment error rate of about 38% on a blind test set.

✍ **4.** The authors use the alignment error rate (**AER**) on the development data normalized so all weights sum to one.

✍ **5.** The authors also confirmed previous findings about the robustness of **SMT** to word alignment. The gain from improving word-alignment quality does not transfer to translation quality. In this case, an improvement actually seems to hurt performance, perhaps because the approach diverges from a purely statistical approach. The results indicate that **AER** is the wrong metric to optimize toward, when the purpose of the word alignment is as an information source for machine translation.

Chapter 6

Linguistically Enriched Word-Sequence Kernels for Discriminative Language Modeling

Pierre Mahe´ and Nicola Cancedda

✍ **1. Language modeling** consists in estimating a probability distribution over the sentences (actually, sequences of words) of a language. This process is central to statistical machine translation (SMT), initially formulated following the noisy-channel model (Brown et al., 1993), in which the probability p(t|s) of observing a sentence t in the target language conditionally on a sentence s in the source language is expressed as

$$p(t|s) \propto p(s|t)p(t)$$

✍ **2.** This decouples the modeling problem into:

- ✵ estimating a translation model p(s|t) to quantify how well t conveys the information contained in the source sentence s,
- 🗐 estimating a (target) language model p(t) to assess the likelihood of t as a sentence in the target language.

✍ **3. Sequence kernels** derive a measure of similarity between sequences by means of their common subsequences.

✍ **4.** Our future work will be mainly dedicated to the actual integration of these techniques in SMT systems. A rapid way to assess their impact, which does not require applying complex modifications to the decoder, is to adopt a **reranking** approach. **Reranking** casts translation into a two-step process.

- 🗎 To translate a given sentence, the **first step** is to produce an "**n-best list**" of candidate translations by the decoder: these are the n top-ranked translations according to the log-linear model.
- ❄ In a **second step**, this list of candidates is **reranked** to find a better candidate than the one returned by default by the decoder (that is, the first one in the **n-best list**).

✍ **5.** When informative features not directly accessible by the decoder are used for **reranking**, this approach can improve the **fluency** of the produced sentences, and is now a standard component of **SMT** systems.

✍ **6.** The mainstream approach to learn **reranking** models is to use perception algorithms trained from a development set, kept aside from training, according to automatic criteria such as the BLEU or the NIST scores. There are at least two simple ways to integrate kernels in such models:

- ❁ Perceptrons being straightforward to "**kernelize**," a first approach would be to integrate directly the kernels in their scoring functions. This would be very similar to the approach presented in Roark et al. (2004) based on n-gram features, which proved effective in the context of speech recognition.
- ☆ An alternative approach, in the direct continuity of this work, would be to first train an **SVM** model to distinguish between fluent and disfluent sentences, and to use the resulting scoring function as a single additional feature to learn the reranking model.

PART II

Machine Translation

Chapter 7

Toward Purely Discriminative Training for Tree-Structured Translation Models

Benjamin Wellington, Joseph Turian, and I. Dan Melamed

✍ **1.** Discriminative training methods have recently led to significant advances in the state of the art of machine translation (**MT**). Another promising trend is the incorporation of syntactic information into **MT**

systems. Combining these trends is difficult for reasons of system **complexity** and **computational** complexity.

✍**2.** The authors' main innovation is an approach to discriminative learning that is computationally efficient enough for large statistical **MT** systems, yet whose accuracy on translation subtasks is near the state of the art. The authors' approach to predicting a translation string is to predict its parse tree, and then read the string off the tree. Predicting a target tree given a source tree is equivalent to predicting a synchronous tree (*bitree*) that is consistent with the source tree.

✍**3.** The authors' method for **training tree transducers** was to train an inference engine to predict *bitrees*. The inference engine employs the traditional **AI** technique of predicting a structure by searching over possible sequences of inferences, where each inference predicts a part of the eventual structure. Thus, to train a model for predicting *bitrees*, it is sufficient to train it to predict correct inferences. However, unlike most approaches employed in natural language processing (**NLP**), the proposed method makes no independence assumptions: the function that evaluates each inference can use arbitrary information not only from the input but also from all previous inferences.

Chapter 8

Reranking for Large-Scale Statistical Machine Translation

Kenji Yamada and Ion Muslea

✍**1. Statistical machine translation** systems conduct a nonexhaustive search of the (extremely large) space of all possible translations by keeping a list of the current n-best candidates. In practice, it was observed that the ranking of the candidates within the **n-best list** can be fairly poor, which means that the system is unable to return the best of the available **N** translations. In this chapter the authors propose a novel algorithm for **reranking** these **n-best** candidates.

✍**2.** The authors' approach was successfully applied to large-scale, state-of-the-art commercial systems that are trained on up to three orders of magnitude more data than previously reported in **reranking** studies. In order to reach this goal, the authors create an ensemble of **rerankers** that are trained in parallel, each of them using just a fraction of the available data. Their **empirical evaluation** on two mature language pairs, Chinese-English and French-English, shows improvements of around 0.5 and 0.2 BLEU on corpora of 80 million and 1.1 billion words, respectively.

Chapter 9

Kernel-Based Machine Translation

Zhuoran Wang and John Shawe-Taylor

✍**1.** In this chapter, the authors introduce a novel machine translation framework based on **kernel** regression techniques. In their model, the translation task is viewed as a string-to-string mapping, for which ridge regression is employed with both source and target sentences embedded into their respective **kernel-induced** feature spaces. Not only does it suggest a more straightforward and flexible way to model the translational equivalence problem, compared to previous probabilistic models that usually require strong assumptions of conditional independences, this method can also be expected to capture much higher-dimensional correspondences between inputs and outputs.

✍**2.** The authors propose scalable training for it based on the blockwise matrix inversion formula, as well as sparse approximations via retrieval-based subset selection techniques. However, because of the **complexities** of **kernel** methods, the contribution of this work is still mainly **conceptual**. The authors report experimental results on a small-scale reduced-domain corpus, to demonstrate the potential advantages of their method when compared with an existing phrase-based log-linear model.

Chapter 10

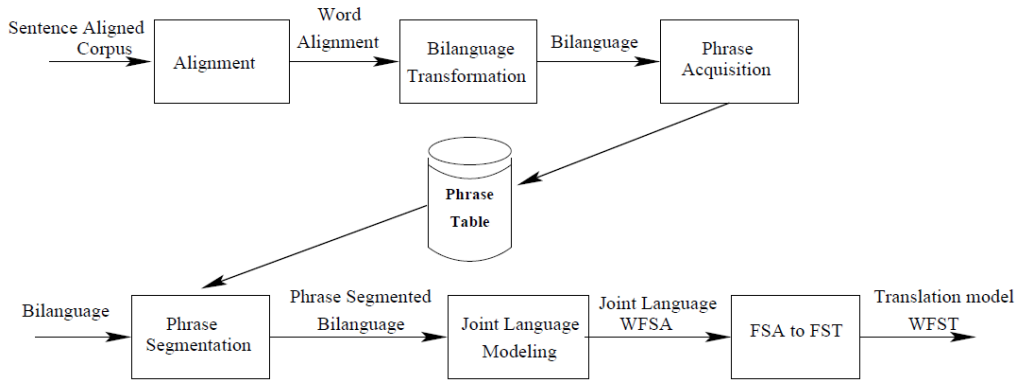
Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction

Srinivas Bangalore, Stephan Kanthak, and Patrick Haffner

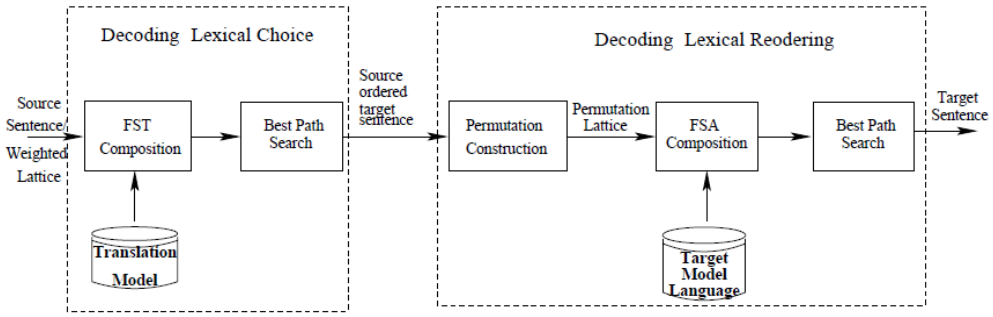
✍**1.** Machine translation of a source language sentence involves **selecting** appropriate target language words and **ordering** the selected words to **produce** a well-formed target language sentence. Most of the previous work on **statistical machine translation** relies on (local) associations of target words/phrases with source words/phrases for lexical selection.

✍**2.** In contrast, in this chapter, the authors present a novel approach to **lexical selection** where the target words are associated with the entire source sentence (global) without the need to compute local associations. Further, they present a technique for **reconstructing** the target language sentence from the selected words. The authors compare the results of this approach against those obtained from a finite-state based statistical machine translation system which relies on local lexical associations.

✍**3. Training phases** for the system:



4. Decoding phases for the system:



Chapter 11

Discriminative Phrase Selection for SMT

Jes´us Gim´enez and Llu´is M`arquez

1. This chapter explores the application of discriminative learning to the problem of **phrase selection** in statistical machine translation. Instead of relying on maximum likelihood estimates for the construction of translation models, the authors suggest using local classifiers which are able to take further advantage of contextual information.

2. Local predictions are softly integrated into a factored phrase-based **statistical machine translation** (MT) system leading to a significantly improved lexical choice, according to a heterogeneous set of metrics operating at different linguistic levels.

3. However, **automatic evaluation** has also revealed that improvements in lexical selection do not necessarily imply an improved sentence grammaticality. This fact evinces that the integration of dedicated discriminative phrase translation models into the **statistical framework** requires further study. Besides, the lack of agreement between metrics

based on different similarity assumptions indicates that more attention should be paid to the role of automatic evaluation in the context of MT system development.

Chapter 12

Semisupervised Learning for Machine Translation

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar

✍ **1. Statistical machine translation** systems are usually trained on large amounts of bilingual text, used to learn a translation model, and also on large amounts of monolingual text in the target language, used to train a language model.

✍ **2.** In this chapter the authors explore the use of **semisupervised methods** for the effective use of monolingual data from the source language in order to improve translation quality. In particular, the authors use monolingual source language data from the same domain as the test set (without directly using the test set itself) and use **semisupervised methods** for model adaptation to the test set domain.

✍ **3.** The authors propose several **algorithms** with this aim, and present the strengths and weaknesses of each one. They present detailed experimental evaluations using French–English and Chinese–English data and show that under some settings translation quality can be improved.

Chapter 13

Learning to Combine Machine Translation Systems

Evgeny Matusov, Gregor Leusch, and Hermann Ney

✍ **1.** This chapter describes how translations produced by multiple machine translation (MT) systems can be combined. The authors present an approach that computes a consensus translation from the outputs of several MT systems for the same sentence. Similarly to the well-established ROVER approach of Fiscus (1997) for combining speech recognition hypotheses, the consensus translation is computed by weighted majority voting on a confusion network.

✍ **2.** Faced with the problem of differences in word order between the system translations, they propose an alignment procedure that learns nonmonotone word correspondences between the individual translations using statistical modeling. The context of a **whole corpus** rather than a single sentence is taken into account in order to achieve high alignment quality.

- **3.** The **confusion networks** which are created from this alignment are rescored with probabilistic features such as system confidence measures and a language model. The **consensus translation** is extracted as the best path from the rescored lattice.
- **4.** The proposed system **combination** approach was evaluated on well-established Chinese-to-English and Arabic-to-English large-vocabulary translation tasks. In their experiments, the authors combined the outputs of five state-of-the-art MT systems. Significant improvements in translation quality in comparison with the best individual MT system have been gained.

Short Answer Items & Tests

5.2 Short Answer Items

1. The early approach to SMT advocated by the IBM group relies on the source-channel approach. This is essentially a framework for combining two models: a word-based ..…….. model and a ..…….. model.

2. In the context of the source-channel approach, the ..…….. model ensures that the output is as grammatical and fluent as possible.

3. NIST is the arithmetic mean of clipped n-gram precisions for different n-gram lengths, also multiplied by a (different) brevity penalty. Also, when computing the NIST score, n-grams are weighted according to their ..…….. .

4. The proposers of METEOR (Banerjee and Lavie, 2005) put more weight on ..…….. than on precision in the harmonic mean, as they observed that this improved correlation with human judgment.

5. Named ..…….. recognition is an important part of many natural language processing tasks. Current approaches often employ machine learning techniques and require supervised data. However, many languages lack such resources.

5.3 Answers

1) translation, language	**2)** language
3) frequency	**4)** recall
5) entity	

☙☆❧ 5.4 Tests ☙☆❧

✍ Select the best choice.

1. The ..... **approach does full analysis and generation, whereas the** ..... **translation approach does a minimum of analysis and generation. The** ..... **approach is somewhere in between.**

a) interlingua, transfer, direct
b) transfer, direct, interlingua
c) interlingua, direct, transfer
d) direct, interlingua, transfer

2. The early approach to ..... **advocated by the IBM group relies on the** ..... **approach.**

a) SMT, source-channel
b) EBMT, source-channel
c) SMT and EBMT, word-based
d) EBMT, transfer-based

3. BLEU is the geometric mean of ..... **n-gram precisions for different n-gram lengths, multiplied by a factor (**..... **penalty) that penalizes producing short sentences containing only highly reliable portions of the translation.**

a) clipped, clarity
b) clipped, brevity
c) blended, accuracy
d) blended, fidelity

4. In the context of the source-channel approach, the ..... **model ensures that the system produces target hypotheses that correspond to the source sentence.**

a) transfer
b) language
c) interlingua
d) translation

5. ..... **forced to include a brevity penalty because** ..... **based only on n-gram precision.**

a) BLEU and NIST are not, they are
b) BLEU and NIST are, they are
c) BLEU (but not NIST) is, it is
d) NIST (but not BLEU) is, it is

☙✷❧ 5.5 Answer key ☙✷❧

	a	b	c	d		a	b	c	d
1			×		2	×			
3		×			4				×
5		×							

Book ❻

Introducing Electronic Text Analysis:

A Practical Guide for Language and Literary Studies

Svenja Adolphs

6.1 Notes

Chapter 1
Introduction

1. To illustrate just some of the kinds of different orientations found in the diverse range of areas that use electronic text analysis, we will consider the examples of **Natural Language Processing** (NLP) and **Humanities Computing** in more detail. NLP is often geared towards developing models for particular applications, such as machine translation software for example.

2. Sinclair (2004b) makes a useful distinction between **description** and **application** in this context.

- Language **description** here refers to the process of exploring corpus data with the aim of developing a better understanding of language in use, while
- an **application** refers to the deployment of language analysis tools with the aim of producing an output that has relevance outside of linguistics.

3. Sinclair (2004b: 55) notes that the end users of **language description** are predominantly other linguists who are interested in empirical explorations of the way in which language is used. The end users of **linguistic applications** on the other hand are not necessarily linguists. They may be people who are simply users of the developed application, such as a spell checker or a machine translation system that has been developed on the basis of a textual resource. The research **goal** in this case is the successful development of an application rather than the comprehensive description of language in use. This distinction marks one of the differences in orientation between corpus linguistics and NLP.

4. A **corpus** tends to be defined as a collection of texts which has been put together for linguistic research with the aim of making statements about a particular language variety.

5. Thorndike (1921) gathered **frequency** information of individual words in a set of texts by manually counting each word form. His **frequency list** was based on a corpus of 4.5 million words from over 40 different sources.

6. A **concordance** is a way of presenting language data to facilitate analysis. The Key Word In Context (**KWIC**) **concordance** has become a standard way of presenting instances of individual lexical items and phrases in a given text or text collection. The search word or keyword appears in the middle of the line with the co-text on either side of the keyword:

on the Friday morning. And I'm	**happy**	to do it with a night sleeper or a
are you	**happy**	guys yeah?
I'm	**happy**	with that. Any fried rice or
Yeah. Certainly. Many	**happy**	returns for tomorrow

7. ***Noam Chomsky*** argued in the 1950s and 1960s that linguistic study should be concerned with the exploration of language **competence**, i.e. the internalized knowledge of a language, rather than language **performance**, the external use of a language (Chomsky 1965).

8. The main argument behind this suggestion was that ***Chomsky*** regarded **performance** data as limited and limiting in terms of what it can reveal about our language competence. He argued that performance can be affected by external events and is thus not an adequate representation of a speaker-listener's language competence.

9. At the same time, ***Chomsky*** noted that **no** collection of naturally occurring discourse can ever be substantial enough to be a true representation of a language.

10. Here is a brief summary of the key **advantages** of using electronic text analysis:

- The reliance on intuition in language research inevitably introduces a high degree of bias into the analysis/description. Using electronic text analysis to study naturally occurring discourse, on the other hand, is a more replicable process and any analysis can be verified by other researchers.
- In addition, electronic text analysis allows us to extract information about language that does not tend to be open to intuitive inspection. This includes information about word frequency and co-occurrence of particular words.
- Electronic text analysis allows us to manipulate language data in various ways to suit a particular research purpose. The use of software tools in this process leads to more accurate and consistent results in a very short amount of time.

- ⌖ Once the data has been sorted in an accessible way, such as in a concordance output for example, we can carry out further analysis on the data. This analysis again helps to identify patterns that we might not be able to describe purely on an intuitive basis. This includes the analysis of whether a word carries positive or negative connotations, and the semantic concepts that surround individual words. It also means that we can identify phrases and clusters of particular types of words.
- ✠ Electronic text analysis can be used at different stages in the analytical process, as required by the researcher. Frequency lists, for example, can give us a good initial overview of our data and further analyses can be carried out on the basis of the derived frequency information. At the same time, we can use electronic text analysis as a hypothesis testing device, where the starting point might be our intuition, which is followed by an analysis of a suitable corpus.
- ☆ Related to the last point is the division between qualitative and quantitative approaches and the direction of progression between the two. Electronic text analysis can be used in a quantitative way, such as through the use of frequency lists, and lead to a subsequent qualitative exploration. Or, it can be used as a secondary method that follows an initial qualitative exploration. An example of the latter approach would be an analysis of frequencies and distributions of a particular language function, such as the use of suggestions in spoken discourse, with the aim of collecting quantitative evidence for results that stem from initial qualitative analyses.

✍ **11. English language teaching.** A large number of dictionaries now include descriptions of words and phrases that are based on corpus research. Other teaching materials, such as grammars and textbooks, are also benefiting more and more from the availability of evidence derived from a corpus. A key advantage of using corpora in this context is that they can provide evidence on word frequencies and distributions in different discourse contexts, which constitutes important information for the language learner. Corpora can also be used as the basis for the description of phrases in a language, which again is of great benefit to the learner. And, finally, there is an increasing body of research that illustrates the discrepancies between the type of English we find in traditional teaching materials that are based on intuition, and the kind that we find in language corpora. As language descriptions evolve with the use of corpora, the integration of new insights into teaching materials seems to be an important next step.

✍ **12. Language variation.** The use of large electronic text collections has facilitated the study of both synchronic and diachronic variation. While

this book focuses mainly on contemporary English, the use of corpora plays an important and extensive part in the study of language development over time. In terms of diachronic variation, the continuous development of new corpora makes it possible to trace language changes over even very short periods of time, and the use of the internet as a resource for linguistic research can help reveal some of the most recent developments in language use. The study of synchronic variation, on the other hand, takes a snapshot of a language at a particular point in time and explores patterns of use in different contexts. The latter has influenced a number of other areas including English language teaching, which benefits from contextsensitive descriptions of language use as they facilitate a more targeted approach.

✍ **13. Language and ideology.** Electronic text analysis has also found an application in the study of ideology (Stubbs 1996). Individual lexical items are being studied in a given corpus with reference to any patterns of usage that show some sort of bias or prejudice. The concept of 'semantic prosodies' developed by Louw (1993) has come to be instrumental in this context. The semantic prosody of a word is the 'shading' of the meaning of that word that can be uncovered through the systematic study of the word in use. The semantic prosody of an individual lexical item may not be immediately apparent through the use of intuition. The study of ideology in language is a particular concern of critical discourse analysts and the use of corpus linguistic methods has opened up a new way of collecting evidence to support theory and practice in this area. Orpin (2005) for example uses a corpus to study words that relate to 'corruption', and finds that words with a negative semantic prosody tend to be used to refer to activities outside of Britain while the same tendency does not apply to words that refer to activities inside Britain.

✍ **14. Forensic linguistics.** Forensic linguistics is concerned with the analysis of texts that are in any way relevant to the law. The particular texts that are studied span a wide range from police interviews to court proceedings (Cotterill 2001). In addition, any other documentation that has legal relevance falls under the remit of forensic linguistics. Electronic text analysis can be used to compare a specific document with a collection of texts where, for example, the aim is to uncover plagiarism or authorship. Forensic linguists sometimes combine corpus linguistic methods with statistics to assess the origin of documents that are relevant in a legal context.

✍ **15. Spoken discourse analysis.** Researchers in the area of spoken discourse analysis, while mainly concerned with detailed descriptions of lexical, grammatical and discoursal patterns in a given stretch of

conversation, have more recently started to draw on multi-million word corpora for their studies. Concordance searches and frequency counts, say for example of discourse markers, often act as a point of entry into the data, as these techniques can highlight particular patterns that can subsequently be subjected to a more qualitative analysis.

16. Sociolinguistics. Sociolinguistics is concerned with the exploration of the relationship between social and linguistic variables. Electronic text analysis has been used to study the occurrence of gender-related language. More recently McEnery (2005) has carried out a large scale corpus-based study of swearing in a number of different discourses. The social variables he considers in relation to bad language range from gender to social class to age, and illustrate the value of electronic text exploration in providing evidence as part of sociolinguistic research.

17. Corpus stylistics. There has been a growing interest in the digitization of literary texts over the last decade. Such resources are often annotated with useful information about the particular text and presented as an integrated archive for the research community. Beyond the ease of access to archives of literary texts themselves and to metadata about the texts, there has been a growing interest in the exploration of literary texts through techniques that have been developed in the area of corpus linguistics. These techniques are applied to organize particular interpretative annotations that have been added to a given literary text or text collection, and can enhance the analysis of literary discourse either in its own right or as a complementary approach that is used alongside other techniques of interpretation. Chapter five will deal with such processes in more detail.

18. Comparing and analysing language varieties. While the main focus of this book is on British English, the development of corpora of other, as well as of more specialized, varieties makes it possible to carry out analyses and comparisons of language use according to regional and national varieties. Recent corpus developments that focus on particular varieties include for example the Scottish Corpus of Texts and Speech (SCOTS), the Limerick Corpus of Irish English (LCIE), and the International Corpus of English (ICE).

Chapter 2
Electronic text resources

1. There are basically **three processes** involved in handling electronic text collections;

- ❁ data collection,
- ✪ annotation and mark-up, and

❋ storage.

✍**2. Mark-up** is the process of adding consistent codes to a text which contain information about its typography and layout. This may include speaker codes in a transcript of spoken data or codes that mark headings or new paragraphs in a written document.

✍**3.** There are various **mark-up systems** currently in use, including **SGML** (Standard Generalized Mark-up Language) and the related **XML** (Extensible Markup Language). These act as a meta-language, which is any language or terminology that is used to describe another language, here used to give additional information about textual features. Both have been adopted by the **TEI** (Text Encoding Initiative), a recognized body that aims to ensure a consistent use of particular coding systems.

✍**4.** Analytical information that is added to a text is often referred to as '**annotation**'. Texts can be annotated automatically by a software program, or in a semi-automated or manual manner depending on the type of annotation that is being used. Annotation is often represented with the use of codes that follow the format of mark-up codes, outlined above.

✍**5.** In terms of literary texts, **annotation** can preserve different interpretations of individual passages or words in a digital format. This can be extremely useful to the research community not only because it aids the preservation of different types of analysis and makes the interpretative processes more explicit, but also because it enables the comparison of different interpretations, either in a manual or sometimes in an automated manner.

✍**6.** In the area of corpus linguistics, the process of **annotation** is closely related to the processes of '**tagging**' or '**parsing**' of texts. The former is a code added to each word in a text and identifies which Part Of Speech (**POS**) individual words represent, while the latter assigns functional categories on the basis of this POS information.

✍**7.** A **POS tagged** corpus allows for a search of lexical items in a particular grammatical role, as well as for a sequence that contains both grammatical categories and lexical items, which will be illustrated further in chapter four. In a tagged corpus, a search of the word play can thus be further specified to include only those instances where play is used as a noun rather than as a verb, as in the concordance lines taken from CANCODE below:

[Nsg] he [Ppers] had [VFpastHave] a [Da] **play** [Nsg] run [Nsg] in [T] the [Dthe] West [Nsg] En
we [Ppers] have [VFpresHave] a [Da] **play** [Nsg] with [T] with [T] the [Dthe] children [Npl]
We [Ppers] did [VFpast] a [Da] **play** [Nsg] scheme [Nsg] with T] the [Dthe] infants
ave] got [VPpast] to [T] a [Da] **play** [Nsg] in [T] the [Dthe] careers [Npl]
[Ppers] saw [VFpast] a [Da] a [Da] **play** [Nsg] but [Cand] obviously [A] because [C] it

✍**8.** A third type of information that is used in the representation of electronic texts is metadata. **Metadata** is '**data about data**' and tends to contain information about the content, source, quality and other characteristics of a particular text. This data can be useful when the corpus is shared and reused by the community and also assists in the preservation of electronic texts. **Metadata** can be kept in a separate database or included as a 'header' at the start of each document (usually encoded though mark-up language).

✍**9.** The last stage in the handling process of electronic texts is that of **storage**, which includes considerations of data access for other users. Most text collections can be stored in the form of a number of different text files in a folder on a standard PC.

✍**10.** Kennedy (1998: 57) makes the following ***distinction*** between **archives** and **corpora**: "In a general sense, databases are collections of information which are designed to facilitate data entry and retrieval. Linguistic corpora, at one extreme, are a subset of databases which have been designed and structured specifically to be used for linguistic description and analysis. Archives, at the other extreme, are usually unstructured repositories of texts."

Chapter 3

Exploring frequencies in texts: basic techniques

✍**1.** The ratio between grammatical and lexical items in the text is referred to as **lexical density**.

✍**2.** A more common ratio, that is often calculated in order to gain some basic understanding of the lexical variation within the text, is the **type-token** ratio.

✍**3.** The term ***tokens*** refers to the number of running words in a text while the term types refers to the number of different words.

✍**4.** The sentence {*This chapter moves from the discussion of design and development of electronic text resources to techniques and practices in data analysis.*} contains 21 tokens and 19 types as the word "**and**" and the word "**of**" occur twice. The **type-token ratio** is calculated by dividing

the number of tokens in a text by the number of types, so the type-token ratio for the above sentence would be 21/19 = 1.11.

✍**5.** This kind of information can be **useful** when assessing the level of complexity of a particular text or text collection, for example in comparisons between documents written for different types of audiences. As a general rule the higher the type-token ratio the less varied the text.

✍**6.** However, the **problem** with the calculation of type-token ratios is that they are dependent on the overall size of the text(s) on which they are based. It is thus advisable only to compare type-token ratios of text(s) of similar length.

✍**7.** Various **word lists** exist in the ELT context, which are based to some degree on word frequency in a corpus, such as the *Academic Word List* (Coxhead 2000) for example.

✍**8. Wordlists** also provide a general picture of a text or collection of texts, and are a good starting point for subsequent searches of individual items at the concordance level. In addition, word lists are useful resources for comparing different corpora, such as those that represent spoken versus written discourse, or American versus British English for example.

✍**9. Wordlists** can be generated by counting the number of identical items in a corpus. This can be done on the basis of **frequency** order, **alphabetical** order, in *lemmatized format* and according to grammatical tags (in corpora that have undergone the POS tagging described in chapter two) and other analytical tags inserted manually or automatically. **Wordlists** can be generated to account for individual items or for recurrent sequences of two or more items.

✍**10. Lemmatized frequency lists** group together words from the same **lemma**, i.e. all grammatical inflections of a word. For example, the words say, said, saying, says are all part of the lemma SAY.

✍**11. Lemmatization** can be done manually using an alphabetical frequency list, or in an automated way which is often based to some degree on lists of predefined lemmas.

✍**12. Frequency lists** can be generated for recurrent strings of sequences, as well as for individual items. The term recurrent continuous sequences describes the consistency of the data string, however, there are a number of other terms in use to refer to such sequences. Biber et al. (1999) use the term '**lexical bundles**' while Scott (1996) refers to them as '**clusters**'.

✍**13. Corpus** research has highlighted the fact that a large proportion of language is **phrasal** in nature, that is, that there is an observable tendency

for particular items to co-occur in a non-random fashion. The attraction between two words is often referred to as collocation.

14. Mike Scott uses the term '**keywords**' to refer to those items that occur either with a significantly higher frequency (positive keywords) or with a significantly lower frequency (negative keywords) in a text or collection of texts, when this is compared to a larger reference corpus (Scott 1997). **Keywords** are identified on the basis of **statistical comparisons** of *word frequency lists* derived from the target corpus and the reference corpus.

15. Key sequences: The analysis of keywords can be extended to include extended recurrent sequences.

16. This chapter has introduced a number of **basic techniques** in electronic text analysis including:

- **The calculation of basic information about a text or collection of texts.** This is an option in a number of concordance packages and includes sentence length, word length, number of paragraphs, ratio between the number of running words in a text and the number of different words in a text (type-token ratio). This information can be used to establish an initial picture of the consistency of text(s).
- **Word lists.** These can be generated in different rank orders including alphabetical, frequency, according to part of speech (POS) and lemma. Wordlists can be used to inform research questions about a text or text collection and to compare two sets of texts or text collections.
- **Keywords and key sequences.** These are words and sequences that occur with a frequency that is significantly higher or significantly lower in a target corpus when it is compared to a larger reference corpus. Keywords and key sequences can be used to profile individual texts and to provide evidence as to the overall orientation of a text.

Chapter 4

Exploring words and phrases in use: basic techniques

1. A **concordance** programme arranges all instances of a particular search item in a way that makes the search item appear in the centre of the page. The search item is also often referred to as the '**node**' and the items to the left and to the right of the node are called the '**span**'. The length of the span can be specified in most programmes but, for descriptive purposes, a span of four or five items to the left and to the right of the node is a commonly used range. In the descriptions of **concordance** data, the **node** is often represented by **N** and the items to the left and to the right as N-1, N-2, etc. and N+1, N+2, etc., respectively.

2. In order to describe the nature of individual units of **meaning**, Sinclair (ibid) suggests **four parameters**: collocation, colligation, semantic preference and semantic prosody.

3. Collocation refers to the habitual co-occurrence of words and will be discussed in more detail below.

4. Drawing on Firth (1957), Sinclair uses the term **colligation** to refer to the co-occurrence of grammatical choices.

5. The **semantic preference** of a lexical item or expression is the semantic grouping of the words that co-occur on either side of the node. In his discussion of the expression *the naked eye*, Sinclair (1996) finds that many of the verbs and adjectives preceding this expression are related to the concept of 'vision'. The verbs 'see' and 'seen' together occur 25 times within four words to the left of the expression in a sample of 151 examples of the naked eye.

6. Sinclair introduces, as his fourth parameter in the description of the unit of meaning, the concept of '**semantic prosody**'. Semantic prosodies are connotations that arise from the co-text of a lexical items and are not easily detected with reference to intuition. **Semantic prosodies** have mainly been described in terms of their positive or negative polarity but also in terms of their association with 'tentativeness/indirectness/face saving' (McCarthy 1998: 22). These four parameters can be used as the basis for a description of words and phrases in a concordance output.

7. Sinclair (1996) argues that there are two principles on which language is based, the '**idiom principle**' and the '**open choice principle**'.

- ❄ The 'idiom principle' operates when speakers make use of lexicalized and semi-lexicalized phrases, which are stored whole in long term memory and retrieved as single items.
- ✰ This principle is opposed to the '**open choice principle**', according to which language is based on grammatical rules and is selected 'slot by slot'. Corpus investigations have shown that a large proportion of discourse is organized according to more or less rigid associations between individual words

8. The term **multi-word units** is used here as an umbrella term for sequences of interrelated words which are retrieved from memory as single lexical units. They occur with varying degrees of fixedness and include formulae (e.g. have a nice day), metaphors (e.g. kick the bucket) and collocations (e.g. rancid butter).

9. Carter (1988: 163) defines **collocation** as, 'an aspect of lexical cohesion, which embraces a "**relationship**" between lexical items that regularly co-occur'. This relationship can be general or text specific, as well as genre

specific. There are various ways in which the attraction between individual lexical items, or in fact between multi-word units and lexical items can be determined. The two techniques are:

- Inspection of concordance data either without further automated analysis or with the help of frequency information.
- Mutual information.

10. One of the **problems** of a **mere frequency calculation** of items that occur within the span of a particular search word is that all of the high frequency, mainly grammatical, items automatically occur at the top of the list. However, since we can expect these to be frequent in the environment of any node item, their cooccurrence is not significant.

11. Mutual information. Apart from deriving collocations through observation of concordance data or through raw frequency information about individual items in the span, there are statistical methods that can be used to account for lexical attraction. Such methods compare the expected frequency with which two words co-occur in a corpus with the actual frequency of co-occurrence. In order to make this calculation, the program calculates the overall frequency of the search word and the individual words of the span in a given corpus. It then calculates the joint frequency of the two, i.e. how often they co-occur. This is referred to as the raw joint frequency. This type of measure does not tell us much about the strength of lexical attraction, since individual words in the span may occur with a very high frequency, as would be the case with grammatical items.

12. The calculation of **mutual information** compares the observed probability of co-occurrence of two items with the expected probability of their co-occurrence. The latter is based on the assumption of random distribution. The ratio between expected and observed frequency is called Mutual Information. The higher this score is, the stronger the attraction between the words.

13. The **advantage** of using **mutual information** over *simple frequency information*: The grammatical items are no longer included in the mutual information output, which makes the analysis of collocates more straightforward.

Chapter 5

The electronic analysis of literary texts

1. Most corpus stylistics studies are designed to either **test** or **facilitate** interpretations of a literary text or collection of texts.

✍ **2.** In terms of electronic explorations of **literary texts**, we can distinguish between two basic types of approaches; those that rely on **intra-textual** analysis and those that are based on **comparisons** of texts with reference to other collections of electronic texts:

- ◈ **Intra-textual** analysis is the manipulation of a text or text collections in a way that might reveal further information about the data, and assist in the interpretation process. This process is particularly useful when we consider longer texts and text collections. The type of manipulation we decide to carry out can be informed by existing interpretations, e.g. a concordance search of words that signal the theme of vagueness, which may have been previously identified, or it can be an exploration of the data that is not guided by previous readings, such as a frequency list of individual words and their collocates for example.
- ▣ Another approach is the **comparison** of individual lexical items and phrases in literary texts with those that occur in other, possibly non-literary, corpora with the aim of analysing deviations and their status as literary effects. This approach might include the analysis of collocations and semantic prosodies, for example. Reference corpora serve as a resource to establish language norms in this context. As such, they can be used as evidence to establish the meaning of individual words and phrases in general language use, which in turn can inform the analysis of such items in a literary text or corpus.

✍ **3.** One of the key differences between **electronic text analysis** and **corpus linguistics** is that a corpus tends to consist of more than one text and, because of its considerable size, it is often impossible to get to know all of its texts in the same way as you would be able to with a single novel for example.

✍ **4.** The study of **inter-textuality** tends to refer to a level of textual reading that takes into account **allusions** to other texts known to the reader, for example religious or historical texts and which thus create a particular literary effect. Corpus linguistic techniques can be used to facilitate **inter-textual analysis** on a number of levels. One way, of course, is to run concordance searches of specific words and phrases from a literary text in the relevant texts and corpora that are **alluded** to, such as the electronic version of the bible in the case of some religious references.

Chapter 6

Electronic text analysis, language and ideology

✍ **1.** The study of **ideology** within the areas of *critical linguistics* and critical discourse analysis focuses on uncovering unequal **relations of power**

through the close analysis of language used to represent certain aspects of society.

✍**2. Ideology** is here used in a negative sense, relating to the goal of **enforcing unequal power relationships**, and pertaining mainly to those types of discourses that pursue this goal.

✍**3.** De Beaugrande (1999) argues that the selection of the types of texts that reproduce a certain **ideology** requires a conceptualization of that ideology that is not based on textual analysis, and thus may, in itself, involve subjective judgement of some kind.

✍**4.** A more **neutral** interpretation of the term **ideology** as a set of beliefs allows for a more inclusive study of different types of discourses, and gets around the problem of subjective choice of texts to some extent. It also allows us to take a broader view of **ideology**, which may include the study of domination and unequal power relationships, but also a more general representation of counter-ideologies, that may become apparent through the study of corpus data (*ibid*).

✍**5.** The focus of this chapter has been on the discussion of the study of **semantic prosodies** as an approach to uncovering attitudes that relate to particular lexical items. The sample analysis has illustrated such an approach with reference to the lemma *GENE*, and has highlighted differences in **prosody** in terms of the different variants of this lemma and the associated discourse in which they occur. This chapter has shown that **corpus-based analyses** of individual lexical items and phrases, that have been identified as relevant references in the study of particular aspects of **ideology**, can be useful in providing evidence from different domains of discourse and from different discourse communities.

Chapter 7
Language teaching applications

✍**1.** As Ellis (1997: 129) argues, 'speaking natively is speaking idiomatically using frequent and familiar collocations, and the job of the language learner is to learn these familiar word sequences'. At the same time, corpus explorations can be carried out by learners themselves and can be used as an integral part of the learning process.

✍**2. Word frequency** information can be used to design syllabuses based on the needs of particular learners with regard to both the sequence of the vocabulary items that are being taught and the overall size of the vocabulary store that is required to achieve an adequate coverage of a language.

While the value of the use of **corpus examples** as part of the material design process is obvious to some, there are a number of interrelated issues attached to this approach that have led to some debate in this area. These issues will be discussed briefly below:

3. The discrepancies between observed language in use and invented examples of language data for materials design. There are now a substantial number of studies that highlight the considerable differences between the language we find in textbooks and the language we find in text corpora. Invented dialogues and multi-party conversations in textbooks are particularly contrived and the common features of naturally occurring interaction, such as ellipsis, turn overlaps, false starts and repetition, are often missing from the textbook data. Similarly, vague language is a particular feature of unplanned discourse which is not generally found in classroom discourse. Sinclair and Renouf (1988) discuss the prevalence of **delexicalized** verbs, such as make and take, in a corpus, and highlight the lack of discussion of such functions in **EFL** textbooks. As outlined in chapter four, one of the main discoveries to come out of the large-scale study of electronic texts is the interrelationship between lexis and grammar. This insight indicates that, in the **ELT** context, syntactic and grammatical structures should not be taught in isolation from vocabulary items, but that we should, instead, be looking to the development of a corpus-driven lexical syllabus in order to reflect the reality of language in use.

4. The tension between pedagogical needs in the ELT classroom and corpus evidence. The question of whether it is desirable for language learning tasks to reflect naturally occurring language use has been raised repeatedly. Language learning tasks are designed to promote particular skills in a specific sequence, which are carefully brought together to reflect a given stage of the learning process. **Invented dialogues** used in textbooks might focus on the acquisition of a specific set of vocabulary, grammatical constructions or speech acts. And it could be argued that the added features of naturally occurring dialogues, such as ellipsis and false starts, for example, might distract from the set learning outcomes. Similarly, we may find that certain naturally occurring dialogues present the learner with vocabulary and grammar that is too advanced for their particular level, where **invented conversations** would allow the materials designers to control the level of vocabulary and grammar knowledge that is required to follow the interaction. On the other hand, it could be argued that every learner should be given the opportunity to engage with the type of language that they are likely to encounter when they are operating outside of the classroom in an L_2, and that we should therefore expose learners to naturally occurring language data wherever

possible. One way of addressing this issue would be to pre-edit corpus data to fit the needs of a particular group of students. However, this would change the nature of the data and run counter to the objective of exposure to naturally occurring language in use.

- **5. The status of corpus data as samples of 'authentic' language.** This issue is closely related to the last one and centres around the notion of authenticity. It is often argued that students should be exposed to authentic samples of language either to facilitate awareness of language in use or to teach particular features of such language samples explicitly. Authenticity refers here to the status of the texts as forms of discourse which have been produced independent of the learning task, in an authentic context, for a particular audience (which tends to be different to that of the language learner). However, some argue that the notion of authenticity relates to the relationship between a particular piece of discourse and the response that it triggers in its immediate audience. This would imply that once texts are taken out of their immediate context, stored in large electronic databases, and reproduced for the teaching context, they are taken out of their authentic environment. The learner, then, has to process such texts with reference to a different context than the one in which they originated, a context which may not reflect his or her communicative goals in the classroom context. On the other hand, naturally occurring data can be contextualized for the learner and the use of such data in the classroom, therefore, not only allows for an integration of some discussion of cultural background to the data, but also empowers the learner by giving him or her the opportunity to engage with genuine language in use.
- **6. The model of the 'native speaker' in the classroom.** This issue is again closely related to the preceding ones in that it relates to the nature of the data that we might wish to include in language teaching material. Recent research has shown that language learners regard the approximation to native speaker English as a main goal in the language learning process. This observation has prompted further exploration of two questions in this debate. Firstly, what do we mean by 'native speaker', and secondly, what is the value of the native speaker model in the **ELT** context. While the notion of the 'native speaker' of English tends to be used to refer to those speakers whose first language is English, this notion is far from unified and remains largely unanalyzed. The vast number of different varieties of 'native speaker' English means that this notion cannot easily be translated into one particular standard for the language classroom. The choice of a particular variety of English for the **ELT** context, even down to fine-grained choices of regional or local variety, becomes a highly political issue. This is, of course, not merely an

issue that relates to naturally occurring data but also to invented examples. At the same time, the proportion of English discourse exchanged between nonnative speakers is growing rapidly, with an overall increase in globalization and internationalization. This raises the question whether native speaker models are the most appropriate basis for language learners, who may predominantly use their L_2 to operate in an international, rather than a 'native', context.

✍**7.** The approach of **data-driven** learning has been developed for use in the ELT classroom (Johns 1991). It is akin to the idea of consciousness raising (Ellis 1993) in the way that it allows the learner to explore language data and thus to **derive patterns** of language use. This approach turns the language learner into the language researcher, giving him or her more autonomy, and by doing so increasing learner motivation. At the same time, letting the learner explore corpus data helps them develop crucial skills of hypothesis testing and data analysis.

Chapter 8
Further fields of application

✍**1.** Tribble (1997) argues that **small corpora** of around 30,000 words can be **useful** to increase language **awareness** of learners and he gives a number of examples of how this may be achieved. The particular value of using corpus techniques in this context is thus related to the successful achievement of the goal, which, in this case, is to raise language awareness.

✍**2. Electronic text analysis** can be valuable in the analysis of **cultural aspects** of language. This is done via a comparison of different language varieties

✍**3. Culture** and **language** are inextricably intertwined and the study of empirical language data should be able to provide some evidence of this relationship. Traces of **culture** in language can be both of lexico-grammatical and pragmatic nature. As such, we may find differing levels of indirectness in the realization of individual speech acts in different cultures for example.

✍**4.** One of the main aims pursued in the area of corpus linguistics is the **identification of language patterns** with a view to establish better language descriptions. This, however, is not necessarily the main aim in other disciplines and in other areas of applied linguistics. The value of particular methodologies and types of electronic text collections therefore has to be assessed in relation to the overall research aim and other methods that might be applied as part of the overall investigation.

5. Three further areas in which **electronic text analysis** can be a useful approach have been discussed in this chapter: discourse analysis, pragmatics and the analysis of language and culture. The sample studies in this chapter have shown how electronic text analysis might be applied in these areas and how it might interface with other, more traditional methodologies. In disciplines that deal predominantly with the **qualitative analysis** of spoken discourse, **electronic text analysis** may be used as an additional source of evidence, which can provide a way in to more **qualitative** types of analyses.

6. The chapter has further listed a range of **challenges** that have to be addressed in order to achieve a more seamless integration of corpus methods in other areas of applied linguistics, humanities and social sciences. These centre around issues of **data collection**, data **representation** and data **replay** and **analysis**, all of which are central to the discussions in the different chapters of this book.

7. The *scope* of electronic text analysis as outlined in this publication then has to be assessed in relation to the **datasets**, **annotation** systems and **computer software** and **hardware** currently available.

Short Answer Items & Tests

6.2 Short Answer Items

1. A ……. tagged corpus allows for a search of lexical items in a particular grammatical role, as well as for a sequence that contains both grammatical categories and lexical items, which will be illustrated further in chapter four.

2. A type of information that is used in the representation of electronic texts is ……. . It is 'data about data' and tends to contain information about the content, source, quality and other characteristics of a particular text.

3. The ratio between grammatical and lexical items in the text is referred to as ……. .

4. Lemmatized ……. lists group together words from the same lemma, i.e. all grammatical inflections of a word.

5. Mike Scott uses the term 'keywords' to refer to those items that occur either with a significantly higher frequency (……. keywords) or with a significantly lower frequency (……. keywords) in a text or collection of texts, when this is compared to a larger reference corpus (Scott 1997).

6.3 Answers

1) POS	**2)** metadata
3) lexical density	**4)** frequency
5) positive, negative	

☙☆❧ 6.4 Tests ☙☆❧

✍ Select the best choice.

1. Noam Chomsky argued in the 1950s and 1960s that linguistic study should be concerned with the exploration of language ……., i.e. the internalized knowledge of a language, rather than language ……., the external use of a language.

a) usage, competence
b) competence, performance
c) performance, competence
d) performance, usage

2. ……. linguistics is concerned with the analysis of texts that are in any way relevant to the law.

a) Legislative
b) Legitimate
c) Court
d) Forensic

3. ……. is the process of adding consistent codes to a text which contain information about its typography and layout. This may include speaker codes in a transcript of spoken data or codes that mark headings or new paragraphs in a written document.

a) Scoring
b) Decoding
c) Mark-up
d) Encoding

4. In the area of corpus linguistics, the process of annotation is closely related to the processes of 'tagging' or 'parsing' of texts. The ……. is a code added to each word in a text and identifies which ……. individual words represent, while the ……. assigns ……. categories on the basis of this ……. information.

a) latter, POS, former, functional, POS
b) former, POS, latter, functional, POS
c) former, function, latter, POS, functional
d) latter, function, former, POS, functional

5. There are various mark-up systems currently in use, including SGML (Standard ……. Mark-up Language) and the related XML (……. Markup Language).

a) Generalized, Extensible
b) Generalized, Explicated
c) Generated, Extensible
d) Generated, Explicated

☙✷❧ 6.5 Answer key ☙✷❧

	a	b	c	d		a	b	c	d
1		×			2				×
3			×		4		×		
5	×								

References

Adolphs, S. (2006). *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies*. New York: Routledge.

Arnold D, Balkan L, Meijer S, Humphreys R. L. and Sadler L. (1996). *Machine Translation: An Introductory Guide.* Colchester: NCC Blackwell Ltd.

Goutte, C., Cancedda, N., Dymetman, M., and Foster, G. (2009). *Learning Machine Translation*. Massachusetts: The MIT Press

Jurafsky, D. and Martin, J. H. (2007). *Speech and Language Processing*. London: Prentice Hall.

Newton, J. (1992). *Computers in translation: A practical appraisal.* New York: Routledge.

Nirenburg, S., Somers, H. and Wilks, Y. (2003). *Readings in Machine Translation.* Massachusetts: The MIT Press.

O'Hagan, M. and Ashworth D. (2002). *Translation-mediated Communication in a Digital World: Facing the Challenges of Globalization and Localization*. Clevedon: Multilingual Matters Ltd.

Quah, C. K. (2006). *Translation and Technology*. New York: Palgrave Macmillan.

Somers, H. (2003). *Computers and Translation: A translator's guide.* Amsterdam & Philadelphia: John Benjamins Publishing Company.

Wilks, Y. (2009). *Machine Translation: Its Scope and Limits*. Sheffield: Springer.

مرور سریع

رایانه و ترجمه (۲)

شامل مهمترین نکاتِ شش کتاب بسیار مفید
درباره‌ی نظریه‌های ترجمه
ویژه دانشجویان کارشناسی ارشد و دکتری
مترجمی زبان انگلیسی و
دانشجویان کامپیوتر و هوش مصنوعی

محمود اردودری
دانشجوی دکتری مترجمی زبان انگلیسی
دانشگاه علامه طباطبائی

دکتر حسین ملانظر
استادیار دانشگاه علامه طباطبائی

۱۳۹۱