In the Name of the Most High

A Rapid Review

Focus on MACHINE TRANSLATION

(1)

Essential Notes and Tests

For M.A. & PhD Candidates

Including:

Translation and Technology (C. K. Quah, 2006)
Computers in translation: A practical appraisal (J. Newton, 1992)
Translation-mediated Communication in a Digital World (O'Hagan 2002)
Readings in Machine Translation (Nirenburg, Somers, & Wilks, 2003)

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 نخستین جلد از مجموعه دو جلدی «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 چهار کتاب مفید در زمینه ترجمه ماشینی است:

- Translation and Technology (C. K. Quah, 2006)
- Translation-mediated Communication in a Digital World: Facing the Challenges of Globalization and Localization (M. O'Hagan & D. Ashworth, 2002)
- Computers in translation: A practical appraisal (J. Newton, 1992)
- Readings in Machine Translation (S. Nirenburg, H. Somers, & Y. Wilks, 2003)

در بخش الحاقی کتاب نیز مجموعه‌ای از مهمترین نکات ذکر شده است. 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

Translation and Technology

C. K. Quah

1.1 Notes

Chapter 1
Definition of Terms

1. This chapter discusses the **definitions** of terms referring to the use of computers in translation activities. Some of the terms can be confusing to anyone who is unfamiliar with translation tools. In some cases, the same **translation tools** are given different names depending on what they are used for; in other cases, a tool may be differently classified depending on the perspective of those who have developed that tool.

2. The aim in this chapter is therefore to **clarify** these **terminological** and related matters. An alternative perspective to the **four basic translation types**–fully automated high-quality machine translation, human-aided machine translation, machine-aided human translation, and human translation–first proposed by **Hutchins and Somers** (1992) is introduced to reflect current developments in translation technology. This will be explored in more detail in the final chapter where the four translation types are reviewed in relation to topics described in the book.

3. *Figure 1.1* distinguishes four **types of translation** relating human and machine involvement in a classification along a linear continuum introduced by Hutchins and Somers (1992: 148).

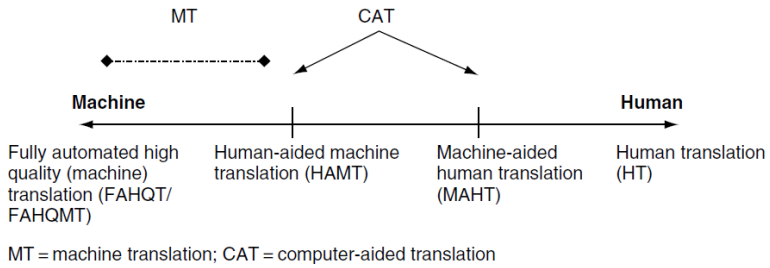


Figure 1.1 Classification of translation types

4. The **initial goal** of machine translation was to build a **fully automatic high-quality** machine translation that did **not** require any human intervention.

5. At a 1952 conference, however, Bar-Hillel reported that building a fully automatic translation system was unrealistic and years later still remained convinced that a **fully automatic high-quality** machine translation system was essentially **unattainable**. Instead, what has emerged in its place is machine translation, placed between FAHQT and HAMT on the continuum of Figure 1.1.

6. The **main aim** of machine translation is still to generate translation **automatically**, but it is no longer required that the output **quality** is high, rather that it is **fit-for-purpose**.

7. In Schadek and Moses (2001), a classification has been proposed where only machine-aided human translation is viewed as synonymous with *computer-aided translation*. Human-aided machine translation is considered as a separate category. For **human-aided machine translation**, the machine is the principal translator, while in machine-aided human translation it is a human.

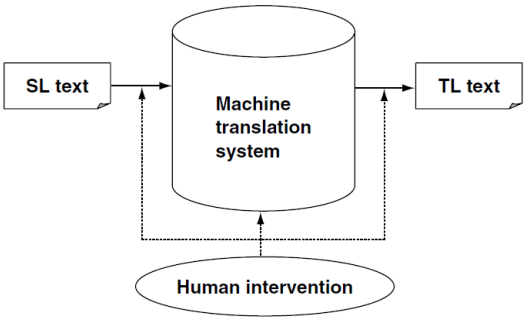
8. The term '**machine translation**' itself can be misleading. The term originally referred only to automatic systems with no human involvement (Sager 1994: 326).

9. The European Association of Machine Translation defines '**machine translation**' as 'the application of computers to the task of translating texts from one natural language to another'.

10. The International Association of Machine Translation (IAMT) defines **machine translation** as taking 'input in the form of full sentences at a time [*sic*] and generating corresponding full sentences (not necessarily of good quality)' (Hutchins 2000).

11. Neither of the definitions above includes **human intervention**. Others, such as Arnold et al. (1994: 1), mention some form of **human intervention**: 'the attempt to automate all *or part of* the process of translating from one human language to another' (my italics). When some form of **human intervention** is mentioned in a definition, it often becomes 'murky' (Balkan 1992: 408).

12. *Figure 1.2* shows how a source-language text can be processed by a **machine translation system**. If the target text is produced automatically there is no **human intervention**; however, human intervention may be employed **before**, **during** and/or **after** machine translation.



SL = source language; TL = target language

13. A machine translation system, according to Hutchins (2000a), can be classified as operating on one of three levels: basic, standard or advanced, each level having its own detailed technical definition given by the IAMT based on the **size of the dictionaries** and the **syntactic analysis** used.

14. A **basic-level** system typically has the following characteristics. It
- has less than 50,000 entries in its largest dictionary,
- has restricted dictionary expansion,
- is restricted to single-clause/basic sentence translations, and
- is suitable for home use.

15. A **standard level** system typically has the following characteristics. It
- has more than 50,000 entries in its largest dictionary,
- allows dictionary expansion,
- allows more than single-clause/basic sentence translations, and
- is suitable for home use and stand-alone office use.

16. An **advanced level** system typically has the following characteristics. It ……
- has more than 75,000 entries in its smallest dictionary,
- allows dictionary expansion,
- allows more than single-clause/basic sentence translations, and
- is suitable for offices with networked facilities.

17. The **size of the dictionaries** and the capabilities of the **syntactic analysis** and synthesis components generally indicate how good a system is.

18. The levels indicated above may **not** necessarily be reflected in **commercial systems** (Hutchins 2000a). An alternative perspective based on **usage** is offered in the compendium compiled by Hutchins, Hartmann and Ito (2004) and shown in Figure 1.3.

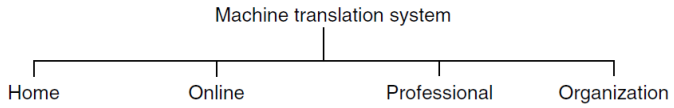


Figure 1.3 Machine translation system based on usage

19. The type labelled '**Home**' refers to machine translation systems for home users who have few or no translation skills.

20. The second type of machine translation labeled '**Online**' is designed specifically for the translation of electronic documents obtained from the Web.

21. The third type is designed for **professional** translators, and the last for **employees** of large companies.

22. A generally accepted view of **human-aided machine translation** is 'a system wherein the computer is responsible for producing the translation per se, but may interact with a human monitor at many stages along the way' (Slocum 1988: 5).

23. In other words, the machine carries out most of the work but it might need human assistance either at the *text-preparation stage* or the **output stage**. The former process is known as '*pre-editing*' and the latter '**post-editing**'.

24. The main task of **pre-editing** is to discover any elements such as **odd** phrases or idioms and typographical **errors** that may create problems for the machine translation system during the translation process. The human editor or translator amends the source language text accordingly.

25. Post-editing involves **correcting** the translation **output** generated by the machine translation system, a task performed by the human editor or translator in order to bring the text to a certain pre-determined standard in terms of language **style** and **appropriate use** of terms.

26. Human intervention is also possible during the translation stage– when prompted by the system–to provide appropriate equivalents for **ambiguous** or **unknown** terms. *Figure 1.4* shows where human intervention [H] is possible.

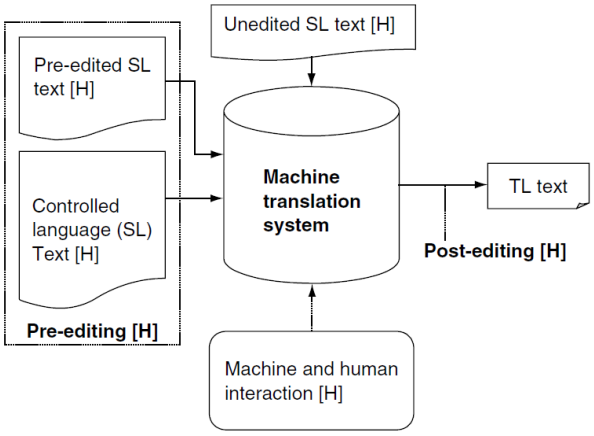


Figure 1.4 Human-aided machine translation model

27. A **source language** text may come in different ***forms***:
- pre-edited
- controlled
- unedited

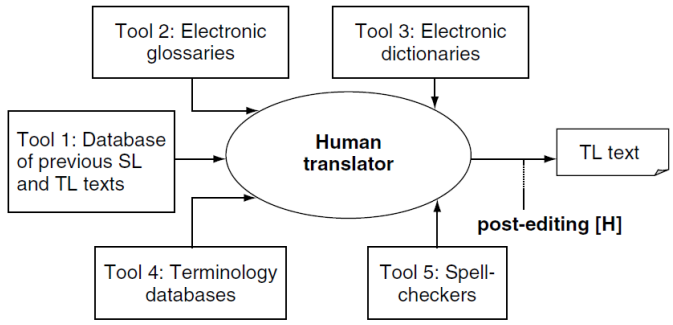
28. A **pre-edited text** is one that has been edited by a **human**, in most cases by someone other than the author, prior to the translation process, whereas a **controlled-language text** is usually written following certain **strict linguistic rules**. Sometimes, a source-language text can also be edited using the controlled-language vocabulary and linguistic rules. Ideally, pre-edited and controlled language texts are *free from* ambiguity and complex sentences.

29. Examples of human-aided machine translation systems are MaTra Pro and Lite developed at the National Centre for Software Technology based in Mumbai, India, that translate from English into Hindi.

30. Machine-aided human translation has been described as the use of **computer software** by translators 'to perform part of the process of translation' (Sager 1994: 326).

31. Integrated machine-aided human translation systems are sometimes known as '**workbenches**' or '**workstations**', as they *combine* a number of *tools*.

32. *Figure 1.5* shows that the ***focus*** in this type of translation is on the **human translator**, who uses an assortment of **tools** such as spell-checkers, electronic glossaries, electronic dictionaries, terminology databases and collections of previously translated texts and their originals, that is translation '**memory**', to support the translation process.



SL = source language; TL = target language; H = human

Figure 1.5 Machine-aided human translation model

33. Some **examples** of commercial machine-aided human translation systems are the Translator's Workbench by Trados GmbH, Transit by Star AG, SDLX Translation Suite by SDL International and Déjà Vu by Atril.

34. In any discussion of **translation technology**, the significant role played by the **localization industry** cannot be ignored. Traditionally, the localization industry has consisted of two sectors:

- the manufacturers of **hardware** and **software**
- the localization **service providers**

35. Localization is the process of **changing** the **documentation** of a product, a product itself or the delivery of services so that they are appropriate and acceptable to the target society and culture.

36. Localization concerns the **changes** required to cater to the **needs** of a particular '**locale**' (Esselink 2000: 3), that is a group of people tied through a shared language and culture.

37. An **example** of the process is the translation and adaptation of ***Time*** magazine into Portuguese and Spanish for Latin American readers.

38. From a **translation** point of view, **localization** is mainly but not entirely a **linguistic** task that involves transferring the text as naturally as possible into the target language, to make the translation '**linguistically** and **culturally** appropriate' for a specific market (Esselink 1998: 2).

39. However, **localization** goes beyond the mere linguistic and adjustments to measurements: target audiences may perceive colours, icons and symbols differently. Thus organizations have to tailor their products to match the **language** and **culture** of the countries they intend to do business in, including countries with ***different varieties*** of the same language.

40. Until the early 1990s, the time when the **Internet** began to be used worldwide, the **translation types** given in Hutchins and Somers (1992) were certainly applicable. More than a decade later, the **boundaries** of these four translation types have become more **blurred**. Although many writers in the field still make clear distinctions, these have become harder to maintain as **technology** becomes increasingly **multifunctional** and more **multitasking**. The pace of change in the development of **translation technology** is extremely **rapid**; what is current today may become outdated tomorrow.

Chapter 2

Translation Studies and Translation Technology

1. This chapter discusses technology within the larger framework of Translation Studies as a discipline, focusing on the relationship between the **engineering of translation technology**, on the one hand, and **Translation Studies including translation theory**, on the other hand. The relationship between academic and professional groups involved in translation is also examined. This in turn leads to a discussion of the involvement of a particular approach in linguistic theories–known as '**formalisms**' in natural-language processing–especially in the design of machine translation systems.

2. A different perspective on the translation process involving **pre-** and **post-editing** tasks using a special variety of language called '**controlled language**' is also presented. This translation process is described using the **translation model** proposed by **Jakobson** (1959/2000), a translation model that **differs** significantly from the one proposed by **Nida** (1969).

3. According to Chesterman (2003), the notion of **translation theory** is '**fuzzy**'. It is also said to be '**a misnomer**, **a blanket term**' (Newmark 1981: 19).

4. A **translation theory** may refer to many different things such as hypotheses, models, assumptions, beliefs, concepts and doctrines. It has numerous interpretations but only **one aim**: to increase the understanding of translation phenomena.

5. Translation theory is in one view an attempt to **create a model** of **how** messages are transferred from a source-language text into a target-language text by giving 'some insight into the relation between thought, meaning and language' (Newmark 1981: 19). It is concerned with **what** is transferred and **why**.

6. *Figure 2.1* shows the approximate **chronological continuum** of translation theories ranging from '**word-for-word** versus **sense-for-sense**' prior to the early twentieth century to a number of different approaches emerging in the 1970s. We see here that translation theories evolved from simple **word-for-word** versus **sense-for-sense** or 'literal versus free' approaches into something considerably more complex.

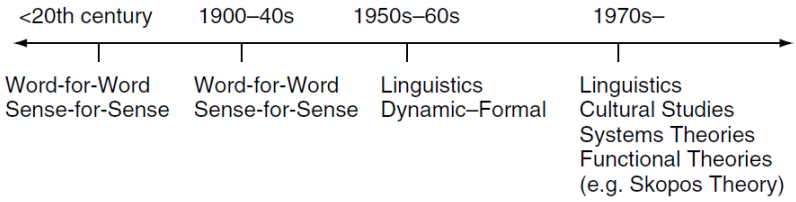


Figure 2.1 Chronology of translation theories

7. The period between the 1950s and the 1960s saw the dominance of **linguistic theories** that focused on the **description** and **analysis** of translation procedures, for example Vinay and Darbelnet (1958/2000), and **typologies of equivalence**, for example Catford (1965).

8. Vinay and Darbelnet's work identifies a number of different **strategies** and **procedures** of translation. Although their analysis was restricted to English and French, the seven procedures that they introduced, ranging from simple **borrowing** of a source-language word into the target language to the more complex procedure of **adapting cultural references** that do not exist in the target-language culture, have had a wider impact.

9. The period between the 1950s and the 1960s saw the **return of the dichotomy** of oppositions similar to that of word-for-word versus sense-for-sense such as '**formal** versus dynamic' as proposed by ***Eugene Nida*** (1964), where the former leans toward the **source-language text structures** while the latter adapts the translation more closely to the target language in order to achieve naturalness.

10. In the late 1970s, another similar dichotomy was introduced by ***Juliane House*** in the form of '**overt** versus covert'. While in '**overt**' translation, it is clear that the target-language text is a translation from another language, 'covert' translation does **not** show that the target text originates in another language.

11. In the early 1980s, ***Peter Newmark*** introduced the dichotomy of '**semantic translation**', which follows as closely as possible the semantic and syntactic structures of the source language text, and 'communicative translation', which is focused on the reader and 'attempts to produce … an effect as close as possible to that obtained on

the readers of the original' (Newmark 1981: 39), recalling Nida's well-known 'dynamic equivalence'.

✍**12.** With *Syntactic Structures* and *Aspects of the Theory of Syntax*– published in 1957 and 1965 respectively–***Noam Chomsky*** introduced '**transformational grammar**' or more specifically '**transformational-generative grammar**' and, as a result, changed the way language could be studied.

✍**13.** Here, the grammar attempts to define **linguistic rules** that can produce an **infinite** number of grammatical sentences in a language from a set of **finite rules** and a lexicon. With his original transformational-generative grammar, Chomsky proposed that a sentence has two levels of representation in the form of an underlying **deep structure** and a surface form which can be mapped onto the semantic deep structure via '**transformations**', such as passivization, pronominalization and topicalization.

✍**14.** In the early days of translation theory, Nida's idea of the translation process as working from the source text to the target text by reaching down to an **underlying level of meaning** as the means of '**transfer**' between the languages resonates with Chomsky's model, as we shall see below in *Figure 2.2*.

✍**15.** The emergence of ***Skopos* theory** is seen as part of a general shift from predominantly **linguistic based** translation theories to a theory that has an orientation towards the way a translation functions in the target **society** and **culture**.

✍**16.** In Greek, ***skopos*** means aim, purpose, goal, objective and intention. It is a technical term used by ***Hans Vermeer*** (1996) to refer to the **purpose of a translation**, which determines the strategy to be used during the translation process (Munday 2001:79).

✍**17.** ***Skopos* theory** allows a source-language text to be translated into a number of *different* target-language texts depending on the **purpose** specified in the so-called 'translation commission' or **brief**.

✍**18.** Vermeer's ***Skopos*** theory draws heavily on the '**translational action theory**' developed by Justa ***Holz-Mänttäri***, which represents a **function-oriented** approach to the theory and practice of translation. A source-language text is an 'offer of information' ('*Informationsangebot*') made by the source language author to his/her recipients. The translation of the source text is then characterized as an '**offering**' of that same information to another **culture** in its own language. The way the translation is *performed* is determined by many factors such as the **needs**, **expectations** and **culture** of the target-language text recipients. Thus,

translation is seen as a process of intercultural communication where the translated text is capable of functioning according to specific target situations and uses (Mason 1998: 33).

19. With respect to **machine translation**, the *intended use* of the target-language text will decide, for example, whether the source-language text gets pre-edited and/or the target-language text gets post-edited in line with quality expectations. In other words, translation is guided by how the target-language text will be used by its intended **readers**.

20. Common ground between **Translation Studies** and **translation technology**–and **machine translation** in particular–may be found within ***functional approaches*** to translation.

21. According to Trujillo (1999: 3), the ***Skopos* theory** of translation strategy, for example, 'arose as a response to the growing need for non-literary translation'. The focus on the **purpose** of the target text in relation to its translation setting resonates with a common definition of translation quality as '**fitness for purpose**'.

22. Theories of translation have been influenced by *different* disciplines and the philosophical backgrounds of translation scholars. Consequently, the definition of a '**translation theory**' depends on the **ideology** subscribed to by the **translation scholar** (Chesterman 2000).

23. According to Halliday (2001: 13), **linguists** have introduced nearly all known **translation theories**.

24. To most linguists, **translation theory** is about

- 'the study of **how** things are' including
  - 'the nature of the translation process and
  - the relation between texts in translation' (Halliday 2001: 13)
- '**why** translations are the way they are' (Mossop 2000: 44)

25. The perspective is **descriptive** in nature (Mossop 2000: 44). The goal is to provide explanations by **describing** linguistic usage as it actually is (Crystal 1993: 100).

26. As viewed by many, **translation** is an *extension* of **language studies** (Neubert 1996: 88) or a *sub-field* of **applied linguistics** (Baker 2001: 47); hence the dependence on linguistics as a descriptive and explanatory discipline is inevitable.

27. To most **professional translators**, on the other hand, **translation theory** is about 'how things ought to be:

- what constitutes good or effective translation and
- what can help to achieve a better or more effective product' (Halliday 2001: 13).

✍**28.** For **translators**, *translation theory* is a '**solution provider**' to problems they encounter during translation (Chesterman 2000). **Professional translators** continue to view *translation theory* from a **prescriptive** perspective expecting theory to take on a **problem-solving** role. Thus the two groups have very different ideas as to what embodies translation theory.

✍**29.** The 'antagonism between "**practicing**" [*sic*][professional] **translators** and "**theorists of translation**"' (Lefevere 1996: 46–50) runs deep because each camp has its own traditions and holds the firm opinion that their method is best.

✍**30.** Newmark (1981: 23–36) states that **translation theorists** are concerned primarily with **meaning** and the **varieties of meaning**. They are also concerned with the **appropriate general method** of translation, every type of translation procedure, specific linguistic problems such as *cultural terms* and *metaphor*, and with ensuring that no linguistic or cultural factor is ignored during the translation process. For **translation theorists**, solving the problems of **professional translators** is a matter of interest only when the approaches they have suggested are involved.

✍**31. Translation theorists** as well as **linguists** often have *little interest* in *providing specific guidelines* to **professional** translators and to translation **trainees**, with a few exceptions such as Malone (1988) and Vinay and Darbelnet (1958/1995). Their research focus is to **describe** and **explain** the processes and products of translation (Fawcett 1997; Chesterman 2000).

✍**32.** Newmark (1981: 36) offers a suggestion about what **translation theory** can do for **professional translators**: it can
- ※ show what is or what may be involved in the translation process
- ⁂ offer general principles and guidelines
- ◈ stop translators from making mistakes

✍**33.** Newmark (1981: 36) cautions that **no** translation theory can turn a *bad* translator into a *good* one.

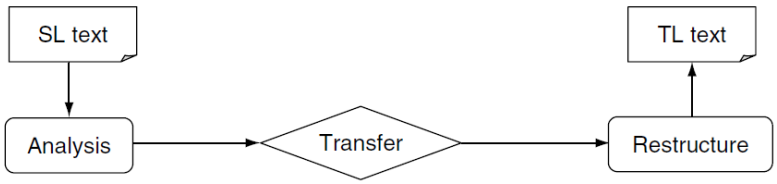
✍**34.** Linguists have been either *neutral* or, at times, even *hostile* to the notion of a **theory of translation** because they have failed to fully understand its objectives and methods.

✍**35.** The notion that **translation** is a '**science**', or perhaps a '**discipline**', is acceptable to *linguists*, who strive to make **objective** observations and descriptions of linguistic phenomena. It is the notion of translation as an '**art**' or '**craft**', as influenced by literary theory and criticism, philosophy and rhetoric, with the creative aspect as the focal point in translation that is a notion less easily embraced by *linguists*, as it is **not** open to objective

description and explanation. As a result of these contradictory views, a theory of translation is often not taken seriously (Bell 1991: 4).

36. Computational scientists have applied **linguistic theories** to enhance the performance of **machine translation** systems because linguistics offers 'a range of observations, techniques and theories that may be adopted and extended within the MT [machine translation] enterprise' (Bennett 2003: 157).

37. Although **rule-based** architectures rely on **linguistic approaches**, they also resemble the three-step translation process introduced by ***Nida*** (1969: 484) as illustrated in Figure 2.2.



SL = source language; TL = target language

Figure 2.2 Translation process model
Source: Nida (1969): 484.

38. For the **scientist**, the main issue is **not** whether linguistics is prescriptive or descriptive; a more important criterion is that the particular approach applied must be **computationally tractable** (Bennett 2003: 144). This means that to be *useful* to the building of a machine translation system, the computer program implementing the linguistic approach must run at a practical or acceptable *speed* on a standard computer.

39. Linguists, on the other hand, are more interested in language from a **human perspective** and since many obstacles are encountered in the process of studying and describing a single language, it is **not** in their interest to even consider studying and describing two languages involving translation. This is compounded by their misconception (like many others) of the real purpose of the development of **machine translation** systems (Hutchins 1979: 29), which is not to replace human translators.

40. Until the late 1960s, the method used to generate translations in nearly all machine translation systems was the '**direct translation**' **approach**. This approach is based on the assumption that **one** target-language word can be generated from **one** source-language word. It also requires a

minimal syntactic analysis, for example, recognition of word classes such as noun and verb (Hutchins 1979: 29)

41. One of the original systems built was the **Georgetown University System**. The poor quality of the translations produced by the system highlighted the complexities of language and the need for a better **analysis** and **synthesis** of texts (Hutchins 1979: 31

42. In subsequent machine translation system designs, two **linguistic approaches** or grammars are considered useful:
- the **formal**
- the **functional**

43. The **formal approach** puts emphasis on the *description* of morphological and syntactic structures.

44. The **functional approach** is concerned with the **use of language** and the ways words and sentences are combined to produce well-formed texts (Bennett 2003: 144).

45. Of the two, the **formal approach** is easier to compute and therefore to incorporate into machine translation than the functional approach, which takes the **pragmatic** view that language is a form of social interaction (Crystal 1993: 146). Thus the **formal approach** has had more influence on machine translation research and development.

46. The focus of the **formal approach** is to establish rules for the formation of grammatical structures: **how** phrases, clauses and sentences are generated (Finch 2000: 99).

47. The **representation method** of formal linguistics involves the conversion of a sentence into a representation that consists of its structure and meaning. A simple example of a representation is illustrated in Figure 2.3:

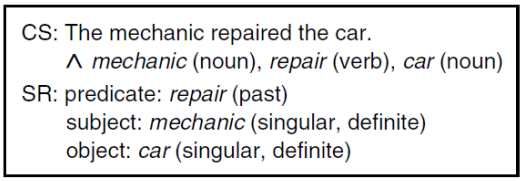


Figure 2.3 Example of sentence representations

48. According to ***Hutchins*, transformational grammar** was found to be **unsuitable** for **machine translation** purposes, as it required extensive and complex computer programming (Hutchins 1979: 33).

49. An alternative approach—arguably better-suited to machine processing—was offered by so-called '**formalisms**', some of which are linguistic **formalisms** while others are found in logic, mathematics and computer science, which can also be applied in machine translation systems. Some formalisms are **syntax-based** such as Chomsky's transformational generative grammar, while others are **lexicon-based**.

50. A **formal grammar** is a set of **rules** that describes a *formal language* (a set of finite words) which is able to represent the syntax of a given sentence. The formal nature of the grammar enables the sentence to be completely analysed by the computer.

51. After the **direct translation** approach, which had much in common with a word-for-word approach to translation owing to the central role of dictionaries in the system, the next generation of systems–known as '**rule-based**' systems–make use of a number of **formal grammars** in the design of machine translation systems.

52. By the mid-1980s, a variant of formal grammar stemming from the '**lexicalist approach**'–different from the *syntax-based* 'constraint-based grammar' or 'unification grammar'–was applied in most **rule-based** machine translation systems.

53. Unification or constraint-based grammar is the general name for a number of linguistic approaches or '**models**':

- '**Tree Adjoining Grammar**', a lexically-oriented grammar that imposes mathematical formalism to capture the syntactic properties of natural languages developed by Joshi, Levy and Takahashi (1975);
- '**Lexical Functional Grammar**', a theory of grammar (syntax, morphology and semantics) by Kaplan and Bresnan (1982);
- '**Generalized Phrase Structure Grammar**', a framework that describes syntax and semantics by Gazdar et al. (1985); and
- '**Head-driven Phrase Structure Grammar**', theoretically influenced by other theories of syntax and semantics, and an immediate successor to Generalized Phrase Structure Grammar, developed by Pollard and Sag (1987).

54. All four models rely heavily on **logic** and **computations** to encode human languages into mathematical codes.

55. The **aim** of unification or constraint-based grammars is to **reduce** the transfer rules – in this case the computational processes of analysis, transfer and synthesis – to simple bilingual lexical equivalences.

56. In Lexical Functional Grammar, a formal description of grammatical units via the 'constituent structure' and 'functional- or feature-structure' is provided (Hutchins and Somers 1992: 39).

✍**57.** The constituent structure or '**c-structure**' consists of groups of phrases analysed as hierarchies. In short, it represents a sentence structure. In c-structure, the rules that identify the grammatical functions are called **phrase structure rules**. An example of the English phrase structure rules for a simple sentence like '*Jane kicks David*' would be:

$S \rightarrow NP\ VP$
$NP \rightarrow N$
$VP \rightarrow V\ NP$
$NP \rightarrow N$

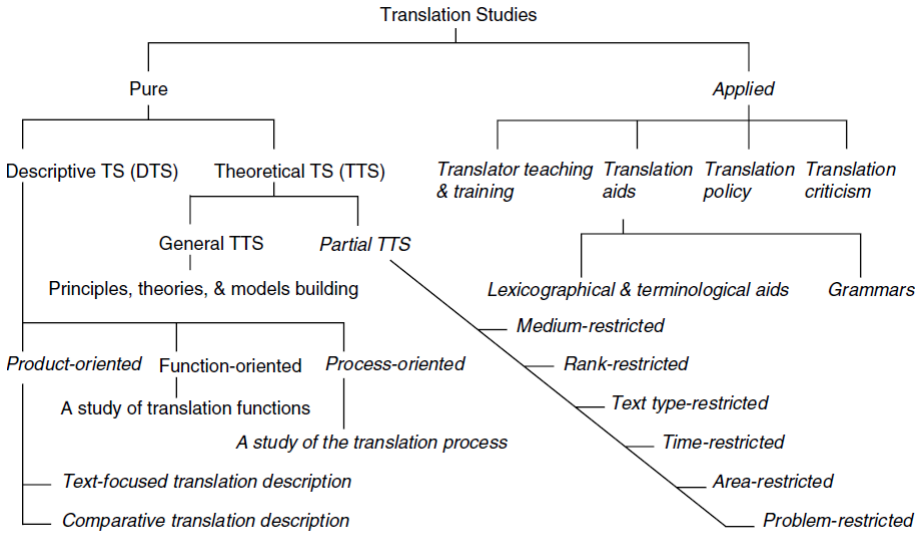
✍**58.** Linguistic researchers such as Peter Toma who founded Systran (System Translation) in 1968 are rare, machine translation research being 'initiated [mostly] by **communications** and **information theoreticians**, and **not** by linguists or TS [Translation Studies] scholars' (Wilss 1999: 141).

✍**59.** As for **professional translators**, there are two possible reasons why they have had little interest in machine translation development (Wilss 1999: 141):

- ❀ First, machine translation is seen as a **distinct area of research**.
- ✠ Second, there is a **lack of knowledge** among many professional translators of programming languages, artificial intelligence (computer programs that can solve problems creatively by making computers behave like humans) and neural networks (systems that simulate intelligence on the computer to imitate the way a human brain works) needed for the development of machine translation systems.
- ⌘ The **absence of translators' input** in the development of machine translation may also be a reason for their resistance to using the technology.

✍**60.** In a seminal paper at the Third International Conference of Applied Linguistics in Copenhagen in 1972, ***John S. Holmes*** put forward a conceptual schema that described various elements of '**Translation Studies**'. It is generally accepted that his paper turned **Translation Studies** into a distinct discipline (Gentzler 1993: 92), now acknowledged as an **interdisciplinary** field (Riccardi 2002b: 2), although it was ***Nida*** (1975) who is widely considered to be the founder of the field of **Translation Studies** as the first to lay down methods of translation in a systematic fashion (Robinson 2003: 13).

✍**61.** *Figure 2.4*: **Holmes' schema** of translation studies:



TS = translation studies; DTS = descriptive translation studies; TTS = theoretical translation studies

62. There are two main branches: **Pure** Translation Studies and Applied Translation Studies. The **Pure** Translation Studies branch has a larger number of levels and sub-branches, consisting at the next level of **Descriptive** Translation Studies and **Theoretical** Translation Studies. The Applied Translation Studies branch of the schema has four sub-branches to do with training, 'aids', policy and translation criticism.

63. Holmes' classification allows the areas of Translation Studies to be seen clearly but it should **not** be taken as '**unidirectional**' as different areas can still influence one another (Holmes 1988/2000: 183). Descriptive Translation Studies, for example, encompasses a host of approaches and disciplines in translation research.

64. The *italicized* branches in Figure 2.4 indicate where in Holmes' scheme we could locate a *strong possibility* of a relationship between **translation** and **technology**, either during the development of various translation systems or at a later stage when they are in use.

65. The objective of ***pure*** research is to describe translation phenomena (**Descriptive Translation Studies**) and to establish the principles (**Theoretical Translation Studies**) that explain these phenomena.

66. Three different types of research are found in **Descriptive Translation Studies**:

- **Product-oriented** research concentrates on the description of existing translations.
- **Function-oriented** research focuses on the description of the impact a translation has on the socio-culture of the target readers.

☆ **Process-oriented** research is concerned with the process of translation itself: what really goes on in the mind of a translator during the translation process?

✍ **67. Think-aloud protocols** (TAP) are one technique used to investigate what comes into the mind of a translator and the actions performed in the creation of a target-language text (Shuttleworth and Cowie 1999: 171).

✍ **68.** Of the three, **product-** and **process-oriented** Descriptive Translation Studies have a higher possibility of **technological involvement**.

✍ **69. Product-oriented** research focuses on existing translations and comes in two forms, text-focused translation description and comparative translation description.

✍ **70. Text-focused** translation description involves describing individual translations of a source text, whereas comparative translation description involves comparing and analysing a number of translations of a single source text.

✍ **71.** In **Theoretical Translation Studies**, the focus is on theoretical work to establish general or partial principles, theories and models.

✍ **72.** The concept of a **'partial' principle** is based on the assumption that a translation theory is limited to researching only certain translation phenomena and can be restricted in more than one way. An **example** of this would be the analysis of novels and short stories written by Gabriel García Márquez, which is restricted to *language* and *culture* (Colombian Spanish into English), *genre* (novels and short stories) and *time* (1960s to the 1990s) (Munday 2001: 12, 192–5).

✍ **73.** The first type is **medium-restricted theories**, referring to the medium that is used to present a text, that is oral (interpreting) and written (translation).

✍ **74.** When Holmes (1988/2000: 178) first described his schema, **interpreting** only involved **humans**, while translation involved humans as well as machines. Nowadays speech technology has developed to the point where it is possible to interpret automatically using machines, thereby opening up new research possibilities in the area of interpreting. It is also possible now for translations/interpretations to be made automatically between written and spoken media

✍ **75.** The second type is **rank-restricted theories**, which is concerned with translation from the point of view of linguistic '**ranks**' *(a Hallidayan term)* or **levels of linguistic analysis**: sentence, clause, group, word and morpheme (Shuttleworth and Cowie 1997: 138).

76. As Holmes points out (1988:179), 'traditionally, a great deal of writing on translation was concerned almost entirely with the **rank of the word'**; this is also reflected in the **direct translation** approach used in first-generation **machine translation** systems, where a word in the source text is matched to an equivalent word in the target text in a kind of '**rank-restricted**' way.

77. The third type is **text type-restricted** theories. The study of text types such as those discussed by Reiss (1977/1989) shows the functional characteristics of three text types and how they can be linked to translation methods.

- ✷ The **informative** type of text ideally uses plain language to convey information, facts and so on in a logical way; examples of informative texts include operating instructions and reports.
- ⁂ The **expressive** type of text uses creative language to express aesthetic form from the author's perspective; examples of expressive texts include poems and plays.
- ※ The **operative** type of text uses a dialogic language to induce desired responses from readers; examples of operative texts include advertisements and sermons (Munday 2001: 73–4).

78. The fourth type concerns **area-restricted theories**, which Holmes interprets as restricted **by language pair** (e.g. translation between French and German) or **language group** (e.g. translation within the Slavic languages), on the one hand, and **by culture** (e.g. within the Swiss culture or between the Swiss and Belgian cultures), on the other hand (Holmes 1988/2000:179).

79. The fifth type is **problem-restricted theories**. This sub-type is concerned with investigating specific linguistic phenomena such as grammatical errors.

80. The sixth type, **time-restricted theories**—which may be focused, according to Holmes, on **contemporary translations** or on translations from an earlier period—could also be developed using electronic corpora and tools. However, **older texts** would need to be converted from paper into digital form, using, for example, an electronic scanner or optical character recognition (OCR). Issues of **copyright** may be relevant here, depending on the date and provenance ofthe texts chosen for analysis.

81. Applied Translation Studies has four subcategories in which the respective objectives of each category are:

- ◈ to improve the quality of translation by developing effective methods of translation teaching and training; to develop better translation tools (or 'aids')

- to establish principles and regulations for professional translators (policy)
- to critique translations

82. Pure Translation Studies research aims for a better understanding of languages, cultures and translation phenomena.

83. Applied Translation Studies can use the information obtained by **Pure** Translation Studies to train translators, to enhance the use of translation tools and to critique translation works (Ulrych 2002: 200).

84. Findings in **Applied** Translation Studies can help researchers of **Pure** Translation Studies to advance their own areas of research.

85. The sub-branch **translation criticism** 'is an essential link between translation **theory** and its **practice'** (Newmark 1988: 184), which can contribute to the development of **translation theories**.

86. Applied Translation Studies is useful for the **localization** industry.

87. With his interest in **Pure** Translation Studies, **Holmes** did not describe applied areas of research in great detail (Munday 2001: 13). However, our concern here is primarily with the **applied** areas, specifically what Holmes calls **translation aids**. Note also that I have replaced the term '**translation aids**' with '**translation technology**' and suggested all the sub-branches below it in order to reflect contemporary developments; these are no longer confined to lexicographical and terminological aids as originally suggested by Holmes (1988/2000: 182).

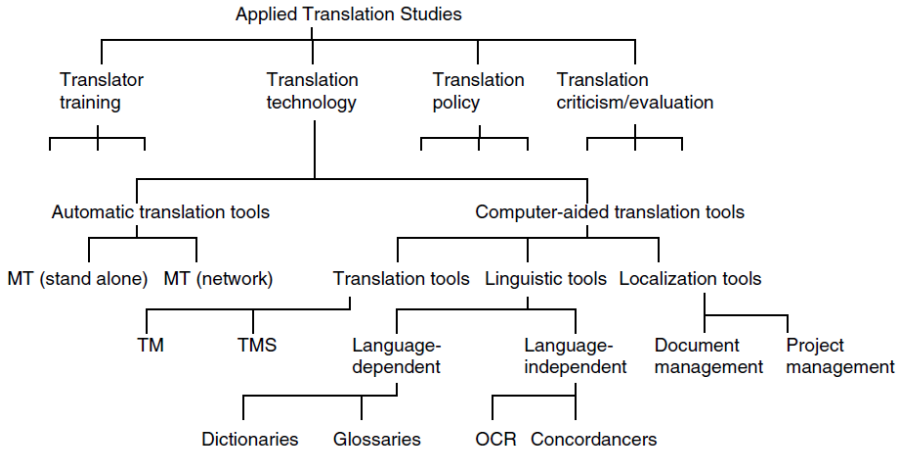


Figure 2.5 A schema of applied translation studies

88. The description of the translation process given below and illustrated in *Figure 2.6* is viewed from a perspective different from those commonly used, for example that of Nida (1969). Here, the tasks of **pre-** and **post-editing** texts as input to and output from **machine translation** systems, and a language variety known as controlled language are described. **Pre-editing** is carried out on a source-language text while **post-editing** is performed on an output (target-language text) generated by a translation tool. **Pre-** and **post-editing** are not always necessary but might be required owing to a number of factors such as the linguistic quality of a source-language text, the type of translation tool used and the required quality of the target-language text (fitness for purpose).

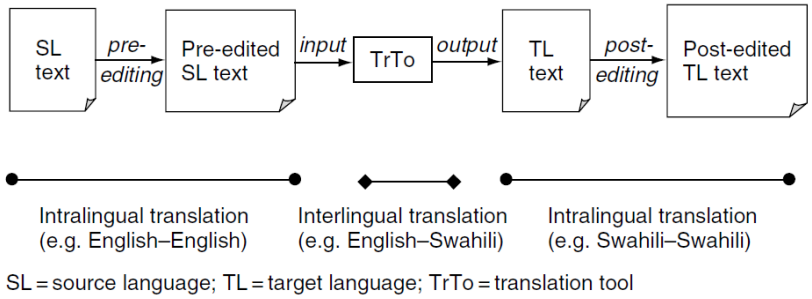


Figure 2.6 A model of the translation process including pre- and post-editing tasks

89. In *Figure 2.6*, Jakobson's (1959/2000: 114) **semiotic categories** of translation are used to characterize the **pre-** and **post-editing** tasks performed when using a translation tool. One of his categories is **intralingual** translation, which is 'an interpretation of verbal signs by means of other signs in the same language' or, in other words, 'rewording', for example, the translation of a poem into prose in the same language (Jakobson 1959/2000: 114).

90. Here we understand **intralingual** translation as **pre-editing** or **post-editing**. The other is interlingual translation, which occurs when a source-language text undergoes a translation process, in this case carried out by a *translation tool* or a *human translator* using a tool, to generate a target text in another language. Interlingual translation is also known as 'translation proper' (Jakobson 1959/2000: 114).

91. *Figure 2.7*: Example of an English SL text and its **pre-edited** version:

SL text in English	Pre-edited SL text in English
Let the water run hot at the sink and then pull the connector from the recess in the back of the dishwasher. Upon the completion of the above task, lift the connector to the faucet by pressing down the thumb release.	1 Turn on the faucet at the sink until the water runs hot. 2 Pull the connector from the recess in the back of the dishwasher. 3 Press down on the thumb release and lift the connector onto the faucet.

SL = source language

✍**92.** Editing a human translation is more commonly referred to as '**revising**'. **Post-editing**, according to ***Laurian*** (1984), is ***not*** a rewriting, revision or correction task but a method of considering a text and working on it for a new aim. ***Allen*** (2003: 297), on the other hand, defines **post-editing** as a task of editing, modifying and/or correcting translated text that a machine translation system has processed and generated.

✍**93. Post-editing** is essentially a specialized skill, which tries 'to preserve as much of the machine's output as possible and "zapping" the text at strategic points rather than redoing it from scratch' (Vasconcellos and Bostad 1992: 68).

✍**94.** A **controlled language** can be defined as 'a subset of a natural language with an artificially restricted vocabulary, grammar and style' (Kaji 1999: 37); one of its **goals** is to improve the *quality* of translation output by humans or machines. It is also employed to restrict the inconsistent use of words and of odd sentences (Wojcik and Hoard 1997: 238).

✍**95.** In other words, the maxim of a **controlled language** is to use simple vocabulary and sentence structures in order to convey complex ideas in writing to ensure rapid reading, understanding, and ease of translation.

✍**96.** A **controlled language** has three important elements: vocabulary, grammar and style (Kaji 1999: 38).

✍**97.** The **size** of the permitted vocabulary is usually restricted to limit the occurrence of lexical ambiguity. The **grammar** restriction occurs at two levels: phrase and sentence. For example, a noun phrase should not consist of more than four nouns, a sentence should not exceed 20 words in length, a paragraph should not exceed six sentences, the passive voice must not be used and the future tense must be avoided (Nyberg, Mitamura and Huijsen 2003: 247). Essentially, a controlled language is **not expressive** and requires some introductory training before a technical writer or professional translator is able to use it.

98. *Figure 2.9*: Example of natural and controlled languages:

Natural language	Controlled language
Remove screws holding the blower and pull the blower from the cabinet. Before the screws are installed to the blower, a new blower is pushed back into the cabinet.	1 Remove screws from the blower. 2 Pull the blower from the cabinet. 3 Push a new blower into the cabinet. 4 Secure the blower with screws.

99. The **best-known** and most widely used **controlled language**, however, is **AECMA** (European Association of Aerospace Industries) Simplified English, a joint effort between AECMA and AIA (Aerospace Industries Association of America). An example of a text for the aerospace industry in AECMA Simple English and its original English text is shown in *Figure 2.10*:

Original English text	AECMA simplified English text
The Model ADI-999 Attitude Indicator (Photo 8) *provides* a *visual display* of pitch and roll attitude and both *enroute Course* Deviation Indicator (CDI)/*Very* High Frequency Omnirange (VOR)/ *Distance* Measuring Equipment (DME)/ *Flight* Management System (FMS) navigation aids and precision approach *Instrument* Landing System (ILS) *information*. The indicator *may be used as* a *long range standby attitude reference*, during a primary power failure, when *coupled with* an emergency power supply. After *complete loss* of power, nine minutes of *useful* attitude *information* is *presented.*	The model ADI-999 Attitude Indicator (see photo 8) has a display for pitch and roll. This display also includes these indicators: • Course Deviation Indicator (CDI) • Very-High Frequency Omnirange (VOR) • Distance Measuring Equipment (DME) • Flight Management System (FMS) • Instrument Landing System (ILS) When there is a power failure, the model ADI-999 can supply an attitude reference for the next nine minutes. After nine minutes, obey the emergency procedures.

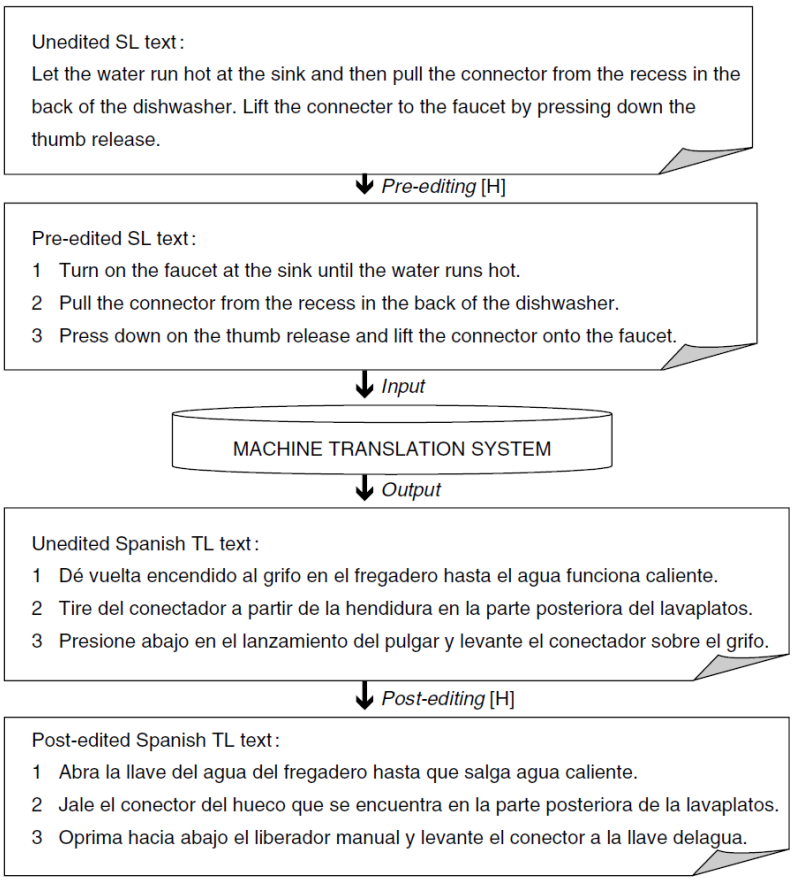
100. Controlled English or indeed any **controlled language** is created by a group of **subject-field specialists** to serve a specific purpose, such as Controlled English for the aerospace industry, a unique variety language which is used **exclusively** in that industry.

101. Controlled English merely selects a specific number of vocabulary items together with their meanings to ensure that **polysemy** and **synonymy** are *eliminated*; as a result, ***one word*** has only ***one sense*** and each sense is conveyed by only one word.

102. Even though **controlled language** commonly occurs in the technical field, *emotional* and *aesthetic* qualities should **not** be excluded entirely

because over-simplification may create other problems. In fact, **controlled languages** should be applied appropriately, pragmatically and sensibly (Janowski 1998).

✍**103.** *Figure 2.13.* Illustration of the **translation process** using a **machine translation system**:



SL = source language; TL = target language; H = human

Chapter 3
Machine Translation Systems

✍**1.** This chapter gives detailed descriptions of different machine translation system designs also known as '**architectures**'. The development of machine translation over several decades, its capabilities and the different types of machine translation systems, past and present, are also included. Both **experimental** and **commercial** systems are discussed, although the focus is on the experimental systems. Even though machine translation

has been well-documented elsewhere, a discussion is deemed to be important for this book.

2. It is felt that modern-day **professional translators** should be informed about machine translation systems because there is every reason to believe that future trends in translation technology are moving towards **integrated systems** where at least one translation tool is combined with another, as is already the case in the *integration* of **machine translation** with **translation memory**.

3. In 1629, Descartes may have been the first to propose the idea that a language could be represented by **codes** and that words of different languages with **equivalent meaning** could share the same code (Pugh 1992: 15).

4. *Figure 3.1* shows the approximation on the time **continuum of approaches** used in machine translation system development since the second half of the twentieth century:

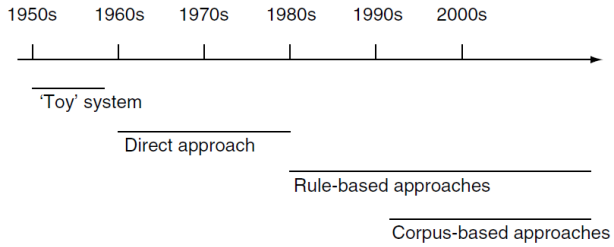


Figure 3.1 Chronology of machine translation development

5. The *first* public demonstration of a machine translation system was the Russian–English **Georgetown University System**, a collaborative effort between IBM and Georgetown University, carried out in 1954 (Hutchins 1995: 434).

6. Early machine translation systems, such as the Georgetown University system, often referred to as the '**first generation**', employed **word-for-word** translation methods with no clear *built-in* linguistic component.

7. Although the Georgetown University system was considered only a '**toy system**' with 250 words, six grammar rules and 49 sentences, it prompted the US government to fund large-scale machine translation research projects (Goshawke, Kelly and Wigg 1987: 26).

8. In 1966, a committee known as the Automatic Language Processing Advisory Committee (**ALPAC**) was established to investigate the feasibility of high-quality machine translation. The **report** concluded that the machine translation systems evaluated were

- slow

✠ less accurate than human translations
☸ expensive

✍ **9.** Thus machine translation systems were deemed a *failure* in meeting their objectives and ALPAC did **not** foresee any possibility of achieving useful results in the near future.

✍ **10.** The **ALPAC report** was considered **biased** due to the unrealistic expectation that machine translation is capable of producing perfect translations of the highest quality. It also did **not** include any study of the long-term needs and possibilities of machine translation systems.

✍ **11.** The **ALPAC report** recommended the development of **machine aids** for translators and shifted its support to research in **computational linguistics**. It also brought about the realization that 'language is too complex and the task of translation therefore requires human capabilities, which . . . cannot be easily simulated in a computer program' (Somers 1997: 194).

✍ **12.** In 1959, ***Bar-Hillel*** argued convincingly that **FAHQMT** (fully automatic high-quality machine translation) should **not** be the goal of machine translation researchers (Nirenburg 1996). He was highly critical of the machine translation projects then in existence, which were mostly **theory-based**.

✍ **13.** The ALPAC report brought machine translation research almost to a **halt**, and as a result the *first half* of the **1970s** was a **quiet period** for machine translation. Some groups continued machine translation research under different names such as 'computer-assisted translation', while others moved on and concentrated on research related to linguistics and artificial intelligence (Tong 1994: 4,731).

✍ **14.** In the **late 1970s**, the USA saw a **revival** of machine translation research with the development of **SPANAM** (Spanish American), a Spanish–English machine translation system, and **ENGSPAN** (English Spanish), an English–Spanish system by PAHO as well as **METAL** (Mechanical Translation and Analysis of Language), a German-English machine translation system built by the US Air Force at the University of Texas in Austin with support from Siemens (Arnold et al. 1994).

✍ **15.** In Europe, between the 1970s and 1992, machine translation research reemerged with the **EUROTRA** (European Translation) project based on the work of the *Groupe d'Étude pour la Traduction Automatique* (GETA) in France and the University of Saarbrücken in Germany.

✍ **16.** In the **1980s**, the most active machine translation research took place in Japan, initiated by the Mu machine translation system developed at Kyoto University.

✍**17.** In the **1980s**, there were also advances in **computational linguistics** that allowed research into machine translation systems to develop more sophisticated approaches to translation. A number of machine translation systems adopted the '**indirect**' approach to translation that was based on certain linguistic rules.

✍**18.** An **indirect approach** enables the source language text to be analysed and turned into abstract representations using programs that can identify word and sentence structures in an attempt to solve the problem of ambiguity. The abstract representations are also able to generate more than one target-language text.

✍**19.** Most of the machine translation systems such as **Pensee** by OKI, **HICATS** (Hitachi Computer Aided Translation System) by Hitachi and **Meltran-J/E** (Japanese/English) by Mitsubishi Electric Corporation are based on the **direct** or **transfer** approach. They all consist of only **word** and **sentence** structure *analysis* with much of the **lexical ambiguities** unresolved. Their domains are restricted to **certain subject fields** such as computer science and information technology. These machine translation systems require **extensive pre-editing** and **post-editing** by human translators

✍**20.** Until the **late 1980s**, two approaches were used in machine translation systems:
- ⌘ the **indirect** approaches
- ❋ the **direct** approaches

✍**21.** The **indirect approach** consisted of two basic systems:
- ▣ interlingua
- ❁ transfer

✍**22.** The best-known **direct machine** translation systems for mainframe computers (a term used to refer to a larger, expensive and more complex computer that processes massive amount of data such as censuses) are Systran, Logos and Atlas.

✍**23.** The best-known **transfer machine** translation systems are Ariane developed by GETA, a machine translation project dating back to the 1960s, and EUTROTRA funded by the Commission of the European Communities.

✍**24.** The **early 1990s** saw another major event when IBM developed a machine translation system called Candide using '**statistical methods**' (Brown *et al.* 1993).

25. At the same time, methods based on corpora of translation examples were experimented with in Japan. This method was later known as the '**example-based**' approach.

26. Neither method, **statistical-** nor **example-based**, used any syntactic or semantic rules, relying instead on large electronic corpora of text to establish patterns of equivalence. Hence, they differ from earlier (prior to 1990) methods such as rule-based approaches that employed linguistic rules.

27. The **statistical-based** machine translation system draws its idea from **communication theory**, which had been suggested nearly six decades earlier by Weaver in his memorandum. In contrast to the **rule-based** approaches, the new **corpus-based** approaches used aligned texts—pairs of source and target-language texts—meaning that the source and target-language texts are structurally matched often at sentence level. **Statistical** calculations are then performed on the **aligned bilingual texts** to establish the probabilities of various translation **equivalents**, or examples are extracted from the aligned bilingual texts by matching examples (strings of source-language and target-language words, phrases or sentences).

28. Since the **early 1990s**, a significant development in machine translation research has been in **speech translation** where speech recognition technology, which deals with the **interpretation** of conversation and dialogue, has combined with machine translation to enable the conversion of speech to text.

29. The achievements of today's machine translation systems are due not only to advances in computer engineering, but also to the realization that in developing such systems, there are limitations. The **limitations** include:

- the **size** of **general** and **specialized dictionaries**
- the **type of text**
- the languages
- the number of language pairs in a system

30. Current machine translation systems are considered to be the **third generation** of hybrid systems that combine the earlier *rule-based* approaches and the subsequent *corpus-based* approaches.

31. A machine translation system normally consists of several main components, and two of these particularly associated with **rule-based** systems are briefly described here. The first component consists of a set of **monolingual** and **bilingual dictionaries**, whilst the second is a **parser**.

32. The function of a source-language **monolingual** dictionary is to present **grammatical** information (morphology, syntax and semantics).

33. A **bilingual** dictionary is consulted by the system when a source-language word is subsequently matched to its target-language equivalent (Lewis 1992: 76).

34. A **parser** assigns a structure to each string made up of a word or phrase in the source-language text based on the stored grammatical information already pre-determined for that language.

35. The **goal** of the **parser** is to identify the relationships between source-language words and their structural representations.

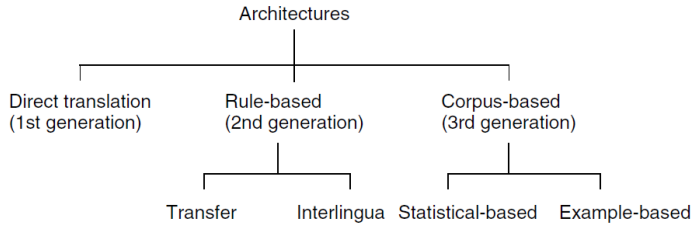
36. A **structural representation** provides grammatical information related to these words or phrases.

37. The word '**supplies**' in the sentence '*The instant hot air **supplies** the necessary heat to all laboratories*' has the structural representations of a verb in the present tense and in the declarative mood (*see Figure 3.2*). The grammatical information is '**attached**' to the words and phrases of the source-language text by means of the *parsing process*.

```
'The instant hot air supplies the necessary heat to all laboratories.'
(*a-supplies
   (tense present)
   (mood declarative)
   (punctuation period)
   (source (*o-hot air
              (reference definite)
              (number singular)
              (attribute (*p-instant))))
   (theme (*u-heat
             (reference definite)
             (number singular)
             (attribute (*p-necessary))))
   (goal_to (*o-all laboratories
              (reference indefinite)
              (number plural)))
```

Figure 3.2 Example of structural representations

38. *Figure 3.3*. Machine translation **architectures**:



✍**39.** The systems from the **second-generation** onwards were designed *differently* from the **first-generation** systems using what is known as a modular structure.

✍**40.** Unlike the second-generation systems, the **direct translation systems** of the first-generation could **not** be modified without the danger of consequent unforeseen changes happening elsewhere in the system.

✍**41.** A **modular approach** means that when grammar rules and dictionaries have to be **updated** or a new pair of languages added, this can be done without affecting the performance of the system as a whole, as the analysis, synthesis, grammar rules and dictionary are separated into different modules.

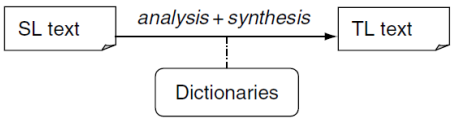
✍**42.** Both later approaches, not only **rule-based** but also **corpus-based** machine translation systems, are **modular**. In all this, it is clear that a machine translation system is not really a machine in a physical sense but a complex software program (Nagao 1989: 70–1, 126).

✍**43.** In **direct** translation systems, **no linguistic analysis** was carried out on the source language text *before* its translation was generated. Also, this approach does **not** have the capability:

✪ to resolve **ambiguities**
✳ to deal with metaphorical expressions
☆ to translate sentences between unrelated language pairs

✍**44.** A **direct** system is essentially a **dictionary-based** system that matches each source-language word to its target-language equivalent. The translation task is a **single processing** operation that stores all data in one *bilingual dictionary* with no separate grammar module (Lewis 1992: 79). The approach mirrors early translation approaches of **word-for-word** translation. It is based on the principle of doing 'simple operations that can be done reliably' and was designed to deal with only one language pair at a time (Jurafsky & Martin 2000: 816).

✍**45.** *Figure 3.4.* **Direct translation model**:



SL = source language; TL = target language

✍**46.** A **direct** translation system depends on:

✷ well-developed dictionaries
⁂ morphological analysis
※ text-processing software

47. Direct translation approach:

- This approach was **simple** and **cheap** but the output results were **poor** and **mimic**—for obvious reasons—the syntactic structures of the source language (Drakos and Moore 2001).
- It only works well with pairs of **closely related languages** that have similar grammatical structures.
- The **syntactic** analysis used is very **basic** while *semantic* analysis is *rarely included*.
- Input to the design of **direct** machine translation systems by linguists and translators was virtually **nil** since this type of system model was designed and built by mathematicians and engineers.
- A **direct** translation approach does **not** incorporate any application of translation theory, and only contains a minimal application of linguistic theory (Somers 1998: 144).
- The machine translation systems resulting from this approach as originally conceived proved to be **unreliable** and insufficiently powerful, yet it was adopted in almost all machine translation systems developed before 1966–67. (Jurafsky & Martin 2000: 817)

48. Ruled-based approaches involve the application of morphological, syntactic and/or semantic rules to the **analysis** of a source-language text and **synthesis** of a target-language text (Carl and Way 2003b: xviii).

49. There are two rule-based approaches: **interlingua** and **transfer**.

50. Rule-based machine translation assumes that translation is a process consisting of analysis and representation of the source-language text '**meaning**' to enable its equivalent to be generated in the target language.

51. As second generation systems, both types of **rule-based** systems have *abstract* or *intermediate* representations:

- The **interlingua** machine translation systems had a **language-independent** or '**universal**' abstract representation, reflecting the aims of theoretical linguists in the 1960s to identify features which all languages have in common at some level.
- **Transfer** systems had **separate representations** for source-language and target-language texts, with the system moving from source-language text to source-language representation, which was then converted into the target-language representation before the target-language representation produced the target text.

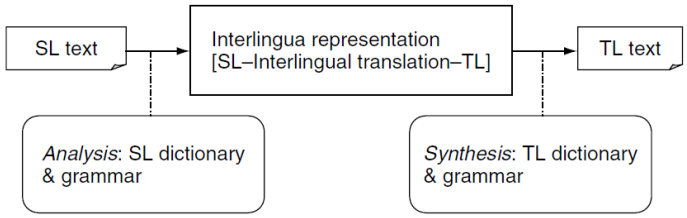
52. In the **interlingua approach**, a source-language text is converted into a highly abstract representation that captures all the essential syntactic and semantic information that can then be converted into several target languages. An '**interlingua**' represents 'all sentences that mean the

"**same**" thing in the same way, regardless of the language they happen to be in' (Jurafsky and Martin 2000: 812). Thus it is designed to be **language-independent**.

53. An **interlingua** is intended to function in stages as the **intermediary** between natural languages.

- During the analysis stage, a source-language text is *analysed* and *transformed* into its *interlingua* representation.
- Target language sentences are produced from this *interlingua* representation with the help of target-language dictionaries and grammar rules during the synthesis stage (Lewis 1992: 78).

54. *Figure 3.5*. **Interlingua model**:

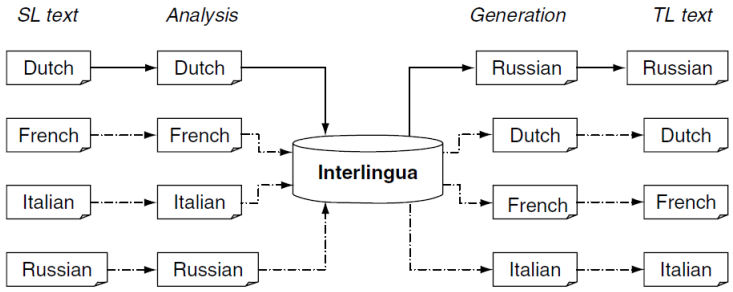


SL = source language; TL = target language

55. Interlingua systems are highly **modular** in the sense that one part of the system does **not** *affect* other parts.

56. Modularity also allows the addition of new modules without affecting existing modules in the system. The **modularity** ensures **independence**; for example, in a Dutch to Russian machine translation system, if the Dutch parser is being upgraded it does not affect the Russian sentence generator.

57. *Figure 3.6* illustrates a **multilingual system** using the **interlingua** approach, which started with Dutch as the source language and Russian as the target language. With **modularity**, it is possible to add three other source languages (French, Italian and Russian) and generate three other target languages (Dutch, Italian and French).



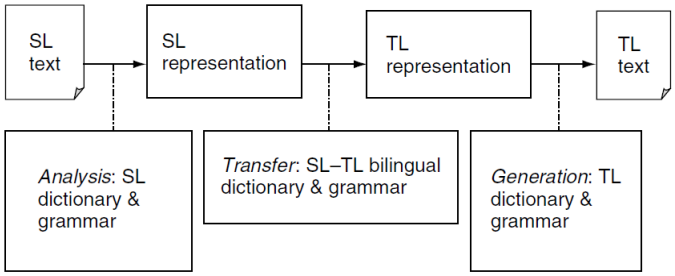
58. The main **problem** for an **interlingua** system to overcome is **how** to define a universal representation that can accommodate **all** languages.

59. The **transfer approach** is *less ambitious* than the interlingua approach, and consists of three stages:

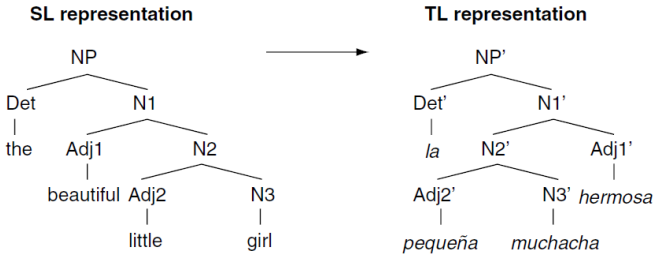
- The **analysis** stage aims to convert a source language text into an abstract source-language representation.
- Following this, the **transfer** of the source-language representation into its equivalent target-language representation takes place.
- The last stage is where a target-language text is **generated**.

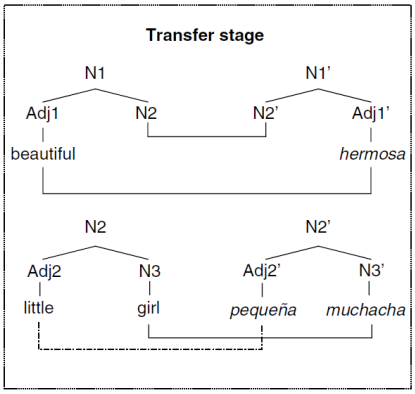
60. The **transfer approach** is similar to the translation process described by ***Nida*** (1969). Specific dictionaries are used at each stage:

- a **source-language** dictionary at the **analysis** stage
- a *bilingual* dictionary at the *transfer* stage
- a target-language dictionary at the generation stage as illustrated in *Figure 3.7*:



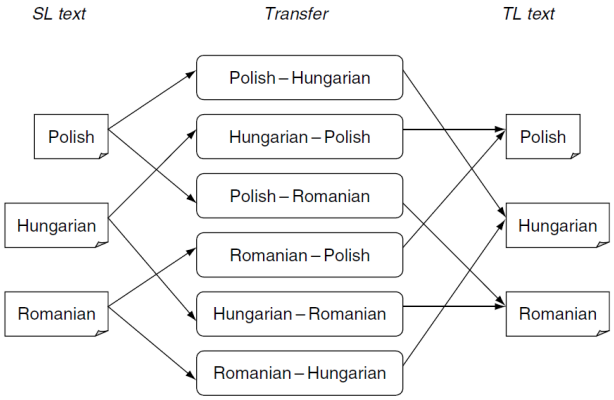
61. The **transfer** approach uses **contrastive knowledge** of the two languages. As an example, *Figure 3.8* shows the transfer stage where the source-language representation of the English phrase '*the beautiful little girl*' undergoes a parsing process to restructure the English phrase into its Spanish translation:





N = noun; NP = noun phrase; Adj = adjective; Det = determiner;
SL = source language; TL = target language
Note: ' = translation (e.g. NP' is the translation of NP).

62. Like the other *rule-based* approach, the **transfer approach** is suitable for building a multilingual machine translation system. However, unlike the interlingua approach where only one interlingua is responsible for all the language pairs, the **transfer approach** uses different transfer models for each language pair. *Figure 3.9* shows an example of a transfer-based multilingual machine translation system of three languages able to generate six language pairs



SL = source language; TL = target language

63. The transfer approach is not without **problems**:

- ☆ It relies on **dictionaries**, which may not necessarily contain sufficient knowledge to resolve ambiguities (Kit, Pan and Webster 2002: 57).
- ✷ Failure at the *analysis* stage may result in **zero** output because the transfer process cannot take place.

64. The ***early 1990s*** saw **corpus-based** approaches gaining popularity in machine translation research. Statistical- and example-based approaches

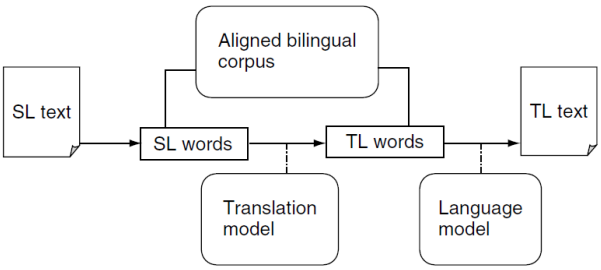
are two different methods that make use of linguistic information in a corpus to create new translations.

65. All **corpus-based** machine translation systems use a set of so-called '**reference translations**' containing source language texts and their translations. Source and target-language texts are **aligned** and the equivalent translation is extracted using a specific statistical method or by matching a number of examples extracted from the corpus (Carl 2000: 997).

66. Corpus-based approaches provide an alternative to the intractable *complexity* of rule-based approaches at the **analysis** and **generation** stages (Hutchins 1994).

67. In the **statistical-based** approach, a source-language text is first segmented into strings of words and phrases; the source-language segments are then compared to an existing large aligned bilingual corpus consisting of original texts and their translations, and a statistical method is then employed on the aligned bilingual corpus to obtain new target-language segments. From the new segments, using the *theorem*, a new target language text is produced (Carl and Way 2003b: xix).

68. *Figure 3.10.* **Statistical-based model**:



69. The principal hypothesis of the **statistical-based** approach is that one source-language sentence (S) can have a large number of translations (T), and each of these has a varying probability (P) of being correct. The probability is calculated using Bayes' rule, which states that:

$$P(T|S) = \frac{P(T) \times P(T|S)}{P(S)}$$

※ where:

P(T|S) is the probability of T given the translation S;

P(T) is the probability of randomly selecting the text T, which is calculated from the frequency in the corpus;

P(S|T) is the probability assumed by the translation model used by the algorithm assigned to S being translated into T;

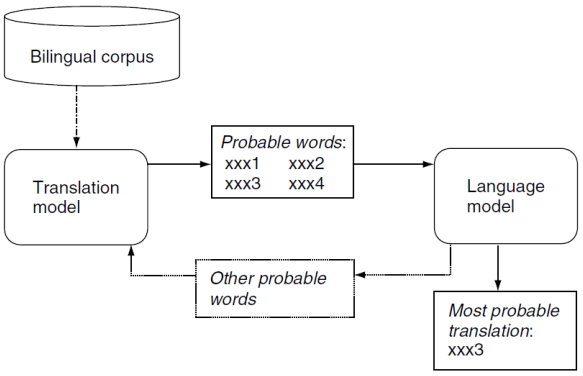
P(S) is the frequency of observing the text S in the corpus.

70. The second component typical of statistical-based systems, the language model itself, is then used to **compute** the likelihood of the results being a valid target-language segment (written as P(S|T) as shown in the Bayes' rules earlier), following the operation of the translation model. The **computation** is best achieved through employing an algorithm, which uses '**n-gram**' statistics.

71. An **n-gram** is a string of 'n' letters. In practice, n is taken as a small number, for example from one to five where n is the number of letters in each of the chosen strings. Therefore, if n = 2 it is called a '**digram**', and if n = 3 it is a '**trigram**'. For example, the text 'the blue car' can be generated using a 'digram' as 'th', 'he', 'eb', 'bl' and so on; or a 'trigram' as 'the', 'heb', 'blu', 'lue' and so on. Note that **n-grams** ignore any spaces between letters.

72. The process of calculating all these **probabilities** can be visualized in *Figure 3.11*. The translation model, as we have seen, is derived from an **aligned bilingual** or **parallel corpus** while the language model calculates the probabilities of word sequences from the target language. Only the most probable translation is usually suggested as the equivalent. Other probable words can also be tried repeatedly to seek better equivalents if necessary.

73. *Figure 3.11*. **Probabilities** *workflow* in the **statistical-based** approach:



74. These *n-gram*-based models **lack contextual information** such as information on the words surrounding the target words, part-of-speech, syntactic constituents and semantics. A **statistical-based** approach also separates the monolingual and bilingual information:

- The monolingual information is located in the **language** model
- The bilingual information comes from the **translation** model (Trujillo 1999: 210–11).

75. A **statistical-based** approach is not without problems. If the *bilingual corpus* is *too small*, the system may **not** be *effective* in generating good translations.

76. Example-based machine translation is also referred to as analogy-, memory-, pattern-, case- or similarity-based translation (Sumita and Imamura 2002).

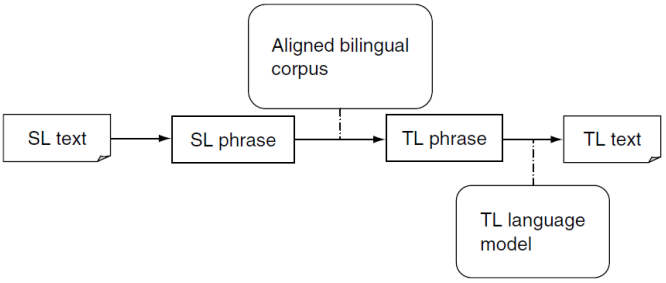
77. ***Nagao*** proposed this approach in the **mid-1980s**, and it lies *between* rule-based and statistical approaches (Carl and Way 2003b: xix).

78. An **example-based** machine translation requires a **bilingual corpus** of translation pairs and employs an algorithm to match the closest example of a source-language segment to its target-language segment as the basis for translating the new source text. A **matched pair** of segments is called an '**example**'. A segment can be of any length or operate at any linguistic level (see also Arnold etal. 1994), but according to one view, **ideally**, it should be at the **sentence level** (Carl and Way 2003b: xix).

79. Three main tasks are involved in the translation process of an **example-based** system:

- **matching** segments from the new source text against existing pairs of examples extracted from an aligned bilingual corpus, then
- **aligning** corresponding translation segments
- **recombining** them to generate a target text (Kit, Pan and Webster 2002: 60)

80. *Figure 3.12.* **Example-based** model:



81. According to Sato and Nagao (1990), the basic idea of an **example-based** translation is to 'translate a source sentence by **imitating** the translation of a similar sentence already in the database'. However, in most cases, more than one 'imitation' may be needed to translate a completely new source-language sentence.

82. The **example-based** approach is very similar to that used in computer-aided translation tools like **translation memory**, which we will consider

in the next chapter. However, while **both** allow translation examples to be extracted from the bilingual corpora stored in the system,

- ❋ Only the example-based approach is capable of extracting more than one example to create a target-language sentence (Trujillo 1999: 203).
- ▣ The other distinction between these two systems is that **translation memory** is an **interactive** tool used by professional translators while example-based machine translation is an automatic translation system (Sumita & Imamura 2002).

✍ **83.** The **example-based** approach is *unlikely to succeed* if **no** close matches can be found in the bilingual corpus or if the input sentences are metaphorical in nature.

✍ **84.** Adding new examples to an aligned bilingual corpus could either improve or degrade the performance of the system. Similarly, too many repetitions of the same or similar examples could either **reinforce** or **jeopardize** the performance of a system. Other areas of concern include how to estimate the **size of the corpus**, and whether the analysis of the corpus should be carried out **before** or **during** the translation process (Sumita and Imamura 2002).

✍ **85. Corpus-based** approaches also have problems with **scalability**, which means that a corpus can be either too small or too large for a particular task (Bel et al. 2001).

✍ **86.** A **rule-based** system is deductive in nature as it is based on a set of linguistic rules set up by its designers. Moreover, it does not in principle store any translation results or reuse previously translated segments. Such a system is difficult to adapt for new subject fields.

✍ **87.** A **corpus-based** system, on the other hand, is inductive in nature because the rules are derived from a given set of translation examples and modification is achieved through the addition of new translation examples.

✍ **88.** The **rule-based** approach is often expensive, and may produce inconsistent results when new linguistic rules are added.

✍ **89.** In contrast, the **corpus-based** approach is flexible enough to process sentences even if they are ill-formed. However, when long sentences are involved, the processing time tends to be lengthy.

✍ **90.** Machine translation research is unlikely to progress significantly by the refinement of **one** approach in preference to another.

- ❁ Instead, '**hybrid**' (Coloumbe 2001) and other innovative approaches may be the best way forward.

❄ Another solution to compensate for the lack of understanding of natural languages on the part of computers is to **involve humans** in the process, that is to have **interactive** machine translation systems.

91. The **human input** or **intervention** feature is similar to that of human-aided machine translation. However, there is one significant ***difference***: interactive machine translation systems allow a translator to have control over the translation process and the output, while human-aided machine translation systems pause and ask the user (not necessarily a translator) to resolve the problem of lexical or syntactic ambiguity. Examples of interactive machine translation systems include LINGSTAT, TransType2 and WebDIPLOMAT.

92. Machine translation developers such as Systran, for example, offer a free **online** translation facility named Babelfish, which is located on the AltaVista search engine website (see http:/ /www.altavista.com/).

Chapter 4
Computer-Aided Translation Tools and Resources

1. This chapter describes the *architectures* and *uses* of several **computer-aided translation tools**, such as translation memory systems, as well as resources such as parallel corpora.

2. Unlike machine translation systems, which are largely developed by universities, most computer-aided translation tools are developed by **commercial companies**. Thus, information about such tools is harder to obtain. This chapter will also show that computer-aided translation tools are becoming more advanced and using different operating systems, and so '**standards for data interchange**' have been created. Three different standards are described. Currently available commercial translation tools are also discussed. In addition, this chapter presents an overview of other commercially available tools such as those used in the localization industry.

3. A '**workbench**' or a '**workstation**' is a single integrated system that is made up of a number of translation tools and resources such as a translation memory, an alignment tool, a tag filter, electronic dictionaries, terminology databases, a terminology management system and spell and grammar-checkers.

4. Translation memory has been defined as 'a multilingual text archive containing (segmented, aligned, parsed and classified) multilingual texts, allowing storage and retrieval of aligned multilingual text segments against various search conditions' (EAGLES 1996).

✍**5.** Unlike machine translation systems, which generate translations automatically, **translation memory** systems allow professional translators to be in charge of the decision-making whether to **accept** or **reject** a term or an equivalent phrase or '**segment**' suggested by the system during the translation process.

✍**6. Translation memory** systems are extremely useful for translating texts that contain large numbers of **repeated** words or terms, extended phrases and even sentences. Legal documents, technical reports and manuals are good examples of texts that can benefit from the use of this type of translation tool.

✍**7.** Generally, a database of terms is known as a '**termbase**'; the tool which is used to build the **termbase** is a database management system which has been customized for storing and retrieving lexical data and is known as a '**terminology management system**'.

✍**8.** A **translation memory** system has no linguistic component, and two different approaches are employed to extract translation segments from the previously stored texts. These are known as **perfect** matching and **fuzzy** matching.

✍**9.** Unlike a perfect match, a **fuzzy match** occurs when an old and a new source-language segment are **similar** but *not exactly identical* (Esselink 1998: 134). Even a very small difference such as punctuation leads to a fuzzy match.

✍**10.** Some translation memory systems are equipped with **filters** for the more common formats. A **filter** is a feature that converts a source language text from one format into another giving a translator the flexibility to work with texts of different formats (Esselink 2000: 362). A **translation-friendly format** contains only written text without any accompanying graphics. In order to obtain such a format, an import **filter** would separate a text from its formatting code.

✍**11.** When the translation is *completed*, the original formatting code can be reincorporated into the translation using the **filter**. The ability to preserve the *format* of a source-language text and apply it to the translation contributes to the robustness of a **translation memory** system (Puntikov 1999: 64).

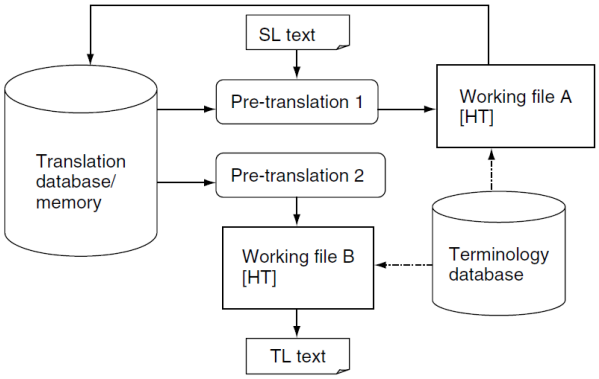
✍**12. Segmentation** is the process of breaking a text up into units consisting of a word or a string of words that is linguistically acceptable. **Segmentation** is needed in order for a translation memory to perform the matching (perfect and fuzzy) process.

13. A pair of old source and target-language texts is usually **segmented** into individual pairs of sentences. However, not all parts of texts, particularly specialist texts, are in a sentence format. Exceptions include headings, lists and bullet points. As a result, **different units** of segmentation are needed. A translator can decide the length of a segment but often **punctuation** is used as an indicator.

14. Alignment is the process of binding a source-language segment to its corresponding target-language segment. The purpose of alignment is to create a new translation memory database or to add to an existing one.

15. The corresponding pairs of source and target-language segments are called '**translation units**'. Once the translator has loaded the **parallel texts** – an original and its translation – into the system, the tool makes a proposal for aligning the segments based on a number of algorithms such as punctuation, numbers, formatting, names and dates, for which the translator is offered various choices. The translator can then **adjust** the alignment proposed by the system before committing the aligned texts to the memory, either by creating a new one, for example for a new subject field or new client, or by adding to an existing one.

16. A typical *workflow* of translation involving a **translation memory** system is described in *Figure 4.3*:

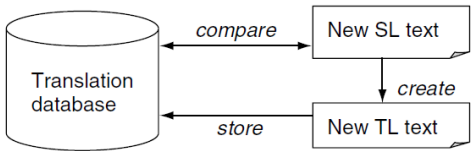


SL = source language; TL = target language; HT = human translator

17. The principal workflow seen in *Figure 4.3* is reflected in almost all translation memory systems, but strategies can follow two models: database and reference (Zerfass 2002). The model shown in *Figure 4.6* has a component that **stores all previously translated** material in one database. The segments are **context-independent**, which allows matching to occur in different translation contexts. Segments from a new source-language text are compared to segments in the database, and translations are offered to the translator if identical and/or similar

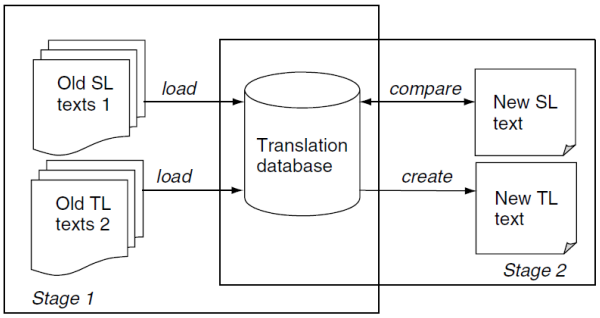
segments are found. Once the translation is completed, a new target-language text is produced and the new or revised segments are added to the database.

18. *Figure 4.6* Database model in translation memory systems:



19. In the reference model, the translation database shown in Figure 4.7 is empty until relevant source and target-language texts are loaded into it in **stage 1**. For example, when translating an updated version of a source language text such as a newer version of an instruction manual, the previous older versions can be aligned and segmented before being loaded into the translation database. Segments from the new source-language text are later compared to the old segments stored in the translation database. Once the translation is complete, a target-language text is created in **stage 2**.

20. *Figure 4.7* Reference model in translation memory systems:



21. For professional translators who specialize in highly technical subject fields, **terminology** is a crucial component of their translation work. A **terminology** – that is a codified collection of terms – can be defined as 'a systematic arrangement of concepts within a special language. **Concepts**, not terms. **Systematic**, not alphabetic' (Bononno 2000: 651).

22. In other words, **terminology** is arranged by concept. Each concept has a label – or set of labels if synonymous – called a 'term', which is a single word or a string of words used to represent it in the language of the specialized field.

23. Concepts are arranged '**systematically**' to reflect the organization of knowledge in a particular subject field, for example to exhibit a hierarchical relationship of scientific classification or taxonomy.

24. A typical **terminology management system** consists of tools to structure the database according to need; a database, which once populated is known as a '**termbase**', and a look-up feature (see Wright and Budin 1997 and 2001).

25. The main functions of a **terminology management system** are to maintain a database, to manipulate terminology resources, to identify multiple equivalents, to establish terminological resources for dictionaries and glossaries, and to exchange terms efficiently (Galinski and Budin 1997: 397).

26. The **database** and **look-up** features are **integrated** in some terminology management systems while in others they are kept separate. **Professional** translators may prefer to use an **integrated** system that enables them to compile a terminology database while translating with a translation memory system. Systems that have separate facilities are more suitable for terminologists. ***Examples*** of commercial terminology management systems are Multiterm by Trados, and Termstar by Star, which can be used separately from their translation memory systems (Translator's Workbench and Transit respectively) while TranslationManager by IBM and SDLX by SDL International are integrated systems (Esselink 2000: 379).

27. A **corpus** in the present context is a collection of **written** texts in a **machine-readable** format.

- In **Translation Studies** and **linguistics**, two terms are used to refer to **corpora** which consist of original texts and their translations: '*parallel corpus*' and '*translation corpus*'.
- In the field of **computational linguistics** the term used is '**parallel texts**'.

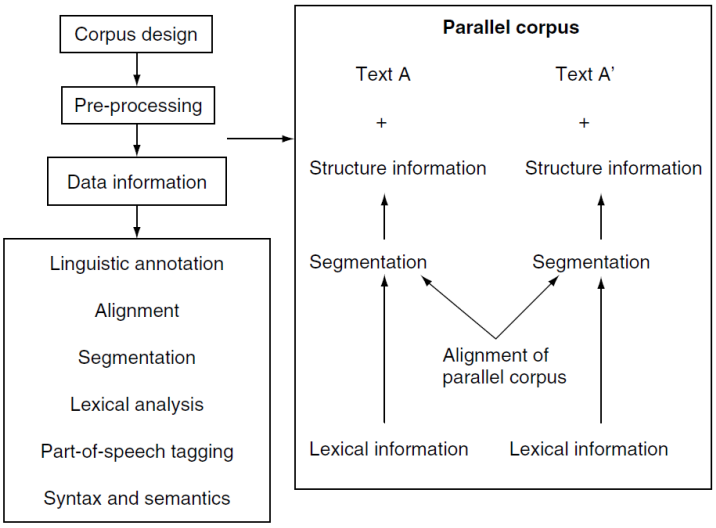
28. Other design possibilities include corpora which consist of texts in **two or more** languages and are selected according to similar predetermined design criteria, for example size, domain, genre and topic. This type of corpus has been called a '**multilingual corpus**' or a '**comparable corpus**' in Translation Studies.

29. Multilingual corpora *cannot be aligned* as there is no source text–target text relationship. However, this type of corpus is rich in useful information for translators (Bowker 2002: 46).

30. The final type is the '**comparable corpus**', which consists of texts in one language, but offering a comparison between original texts and translations into that language.

31. Belonging to the broad field of language technology, **parallel corpora** are used as a linguistic resource for a wide range of applications including the compilation of **termbases**.

32. *Figure 4.8.* Flowchart to illustrate **how to build a parallel corpus**:



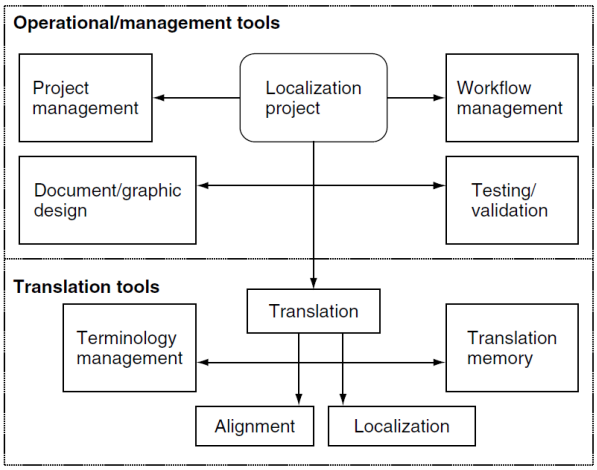
Text A = source-language text; Text A' = target-language text

33. A **concordancer** is an electronic tool which has been used in language learning, literary analysis, corpus linguistics, **terminography** and **lexicography**. It allows the user to select a particular word or phrase and displays the uses of that word or phrase in the selected corpus in order to show where and how often it occurs, and in what linguistic contexts it appears. The output is called a **concordance**. The **concorded** word is shown in the centre of each line displayed in the **concordance**, so that the user can quickly scan the results.

34. While **concordancers** are strictly speaking used to produce concordances, such tools often have other functions, including typically the production of indexes (referenced lists of words from the selected corpus showing where they occur and their frequency distributions) and wordlists, which are like indexes without any indication of text location.

35. Localization tools have been developed in order to support the translation of software applications, product documentation and websites. **Localization tools** are used in conjunction with other computer-aided translation tools such as translation memory systems and terminology management systems.

36. *Figure 4.12.* Types of tool used in a localization project:



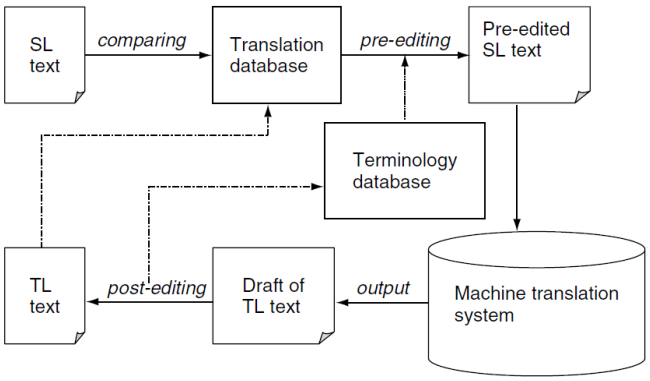
37. A typical **localization process** involves three stages, namely:
- ✠ project preparation
- ✪ the translation proper
- 🗎 quality assurance

38. In the **project preparation** stage, the hardware and software may need to be reconfigured depending on the format of the source-language material, and references related to the subject field of the material may also need to be collected; translators may be required to get training if they are unfamiliar with the subject or with hardware or software applications.

39. In order to prepare the source-language material of the *translation proper*, it undergoes a process called '**localization-enablement**' or '**internationalization**'. This process entails, for instance, stripping all graphics from the text which is to be translated. The purpose is to make it easier to localize and translate a document into a specific language (Esselink 1998: 2).

40. At the stage of the **translation proper**, a translation memory system needs to be prepared by either creating a new database (that is, memory) or using an existing database from another project.

41. *Figure 4.13*. Example of the translation process using a machine translation system, a **translation database** and a **terminology database**:



42. A **standard** is a universal format that has been agreed and approved by either an international standards organization such as ISO or the relevant industry such as the localization industry. In the case of data exchange, the aim of a standard is to facilitate exchange using a common markup language to structure the data in each document using a set of agreed tags as annotations

43. Until recently, most computer-aided translation tools were not compatible with each other, and as a result the import and export of files into and from different software applications presented great problems. In 1998 this prompted OSCAR (Open Standards for Container/Content Allowing Reuse) to create the **translation memory exchange** (**TMX**), an intermediate format to facilitate the sharing of translation memory data.

44. Having **TMX** in a translation memory system increases its flexibility to combine with other computer-aided translation and localization tools. One **goal** of **TMX** is to maximize the reusability of previously translated material, which may have been stored in different formats.

45. *Figure 4.14* illustrates how **TMX** facilitates the sharing of data between different formats. A text in Word format stored in Database 1 is exported to the **TMX** format and then imported into an **HTML** format. The **HTML** text is stored in Database 2. It is important for the import and export processes to apply the same **TMX** specification to prevent the loss of information during the importation and exportation processes:

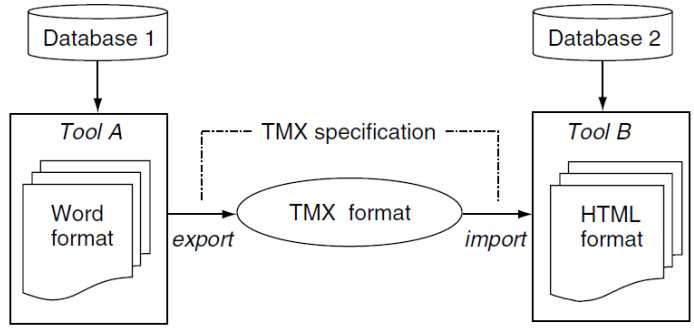


Figure 4.14 Example of TMX data-sharing

46. Between professional translators, sharing terminology is important, and it is also beneficial to anyone wishing to upgrade his or her own terminology databases. This sharing of terminology is called '**terminology interchange**'. Since terminology management systems vary, standards for terminology interchange have been created based on ISO 12620 (Data Categories). Such standards include:

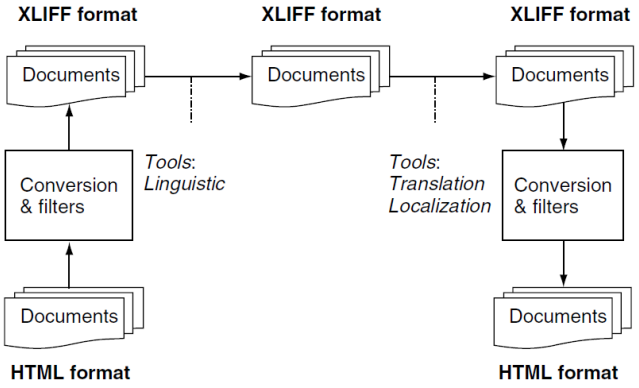
- ☆ **MARTIF** (Machine-Readable Terminology Interchange Format)—(also known as ISO 12200–Computer Applications in Terminology), a format for platform-independent and publicly available terminological data interchange. It functions as a channel for transferring data from one terminology management system to another.
- ✷ **GENETER** (Generic model for Terminology)—a tool to represent terminological data which serves as an intermediate format between different applications and platforms.
- ※ **OLIF** (Open Lexicon Interchange Format)–a tool that exchanges lexical and terminological data. It addresses data management needs for basic terminological exchange and lexicons for machine translation.
- ◈ **XLT** (XML representation of Lexicons and Terminologies)—a standards based family of formats that represents, manipulates and shares terminological data. It is able to merge and extract OLIF, GENETER and MARTIF, and provides the basis for the TermBase eXchange.
- 🗎 **Termado—**a tool that manages and publishes term catalogues, lexicons and dictionaries. It also imports and exports terms to and from external applications such as other terminological standards (MARTIF and OLIF).

47. Texts are often stored in different file formats, some of which are proprietary, for example reports belonging to a company, while others

are commonly shared such as HTML files. These files are not necessarily easily transferable from one tool to another. In order to eliminate such challenges, a standard called **XLIFF** (XML Localisation Interchange File Format) has been developed by OASIS (Organization for the Advancement of Structured Information Standards).

48. XLIFF is another XML-based format that allows the interchange of localization information and is tool-neutral, enabling what is claimed to be a seamless transfer of information between tools (OASIS 2003: 14). The advantage of using **XLIFF** is that it separates a text from its formatting for translation purposes, enables the use of multiple tools and stores information during a localization process.

49. *Figure 4.19.* Example of **XLIFF** in the localization process:



Chapter 5
Evaluating Translation Tools

1. This chapter touches on the **evaluation** of translation technology. The discussion focuses on different groups of stakeholders from research sponsors to end-users. Also included in the discussion are the **different methods of evaluation**: human, machine, and a combination of human and machine as evaluator. The choice of method used depends on who the evaluation is for and its purpose. It also depends on whether an entire tool or only some components are evaluated. Also described in this chapter is the general framework of evaluation offered by various research groups in the USA and Europe.

2. The literature on evaluation concentrates on the evaluation of machine translation systems either **during** the developmental stage or **after** the process of development is completed. Less information is available on the evaluation of computer-aided translation tools. What is available is found mainly in translation journals, magazines and newsletters.

3. The **harshest evaluation** of machine translation came in the infamous **ALPAC report**, which highlighted the misconceptions about language, usage and the system requirements of fully automatic high-quality machine translation systems.

4. For **researchers**, evaluation can reveal if the theories applied yield the desired results; for **developers**, evaluation is a means of showing how good the system is for potential buyers; and for **end-users**, evaluation can provide useful information as to which system best suits their needs

5. In the past, there has been a tendency to concentrate on only **two aspects of evaluation** for machine translation:
- **intelligibility** – the quality of the translation generated by a system
- **fidelity** – the closeness of the translation to its original text

6. For **researchers**, if a system is proven to produce syntactically and lexically well-formed sentences, then such an evaluation may be considered sufficient.

7. For **end-users**, on the other hand, this type of evaluation is often insufficient as other measurements such as coverage (specialization of subject field) and extensibility (the ability to add new words and grammar rules) are equally important.

8. Schmitz (2001) focuses on assessing the criteria for evaluating terminology database management programs, that is terminology management systems. The **criteria** include:
- **terminological** aspects – the suitability of the software to perform a terminological task;
- **technical** aspects – the hardware and software environment required when using a certain tool;
- **user interface** aspects – documentation on how to operate a particular tool;
- **organizational** aspects – compatibility with existing hardware and software; and
- **economic** aspects – purchasing and operating costs.

9. In the evaluation of machine translation systems, there are at least *four groups* with an **interest** in the matter, each with their own set of criteria and goals which may or may not overlap:
- Researchers
- Developers
- Research sponsors
- End-users

10. End-users are made up of several groups, the major ones consisting of translators and translation managers. The evaluation criteria that interest these groups include the 'hows' and the 'whats' (Trujillo 1999: 254).

11. The '**how**' questions include:

- how easy is it to operate a tool;
- how user-friendly is a tool;
- how long does it take to learn;
- how compatible is it with other hardware and software applications;
- how good is the design of the working environment (the layout of the interfaces and display of windows);
- how good is the support for Latin and non-Latin based languages; and
- how easily can a tool be extended or upgraded.

12. The '**what**' questions include:

- what is the processing speed;
- what are the linguistic capabilities;
- what is the required operating system;
- what is the performance reliability; and
- what are the costs and benefits.

13. In the early days, **human evaluators** were used to evaluate translations generated by machine translation systems. *Intelligibility* and *fidelity* are the two main criteria used in evaluation.

14. One example of human evaluators judging the **intelligibility** and **fidelity** of machine translation output relates to a number of early Russian–English machine translation systems evaluated in the ALPAC report.

15. To measure **intelligibility**, 18 English monolinguals were selected to judge six translated texts (three by machine translation systems and three by human translators) using a **scale** from **one**, 'Hopelessly unintelligible . . .', to **nine**, 'Perfectly clear and intelligible . . .' (ALPAC 1966: 68–9). The higher the score, the more **intelligible** the translation.

16. In order to assess **fidelity**, two groups of English native speakers were used. Members of the first group, who were bilingual (English and Russian), were asked to extract *information* from the English translations and compare this with the *information* in the Russian originals. The second group, English monolingual evaluators, were asked to assess the informativeness of the two sets of English translations (one set translated by machines and the other by humans) using a **scale** from **zero**, 'The original contains . . . less *information* than the translation . . .', to **nine**, 'Extremely *informative* . . .' (ALPAC 1966: 68–9). The higher the score, the more *informative* the translation.

17. Sometimes **non-scale** methods can also be used to measure **intelligibility**, for example the **Cloze test** (in which blank spaces at regular intervals must be filled in by the evaluator), **multiple-choice** questionnaires and **knowledge tests**.

18. For **fidelity**, the methods include the **correctness** of the information transferred, **retranslation** and **direct questioning**.

19. A **quantitative evaluation** performed by **machines** is often seen as preferable and is considered to be more stable, reliable and cost-effective.

20. Developers especially are interested in **inexpensive automated evaluation** methods that are ***fast***, language-independent and comparable to evaluations performed by human evaluators.

21. One of the automated evaluation methods that has been designed is **BLEU** (Bilingual Evaluation Understudy), the thinking behind which was that 'the closer a machine translation is to a professional human translation, the better it is' (Papineni et al. 2002).

22. In order to show that **BLEU** is a reliable and objective evaluation method, two criteria were used:

- an evaluation of the '**closeness**' between a translation produced by a machine translation system and a translation translated by a translator, and
- an evaluation of a translation produced by a machine translation system using bilingual and monolingual **human** evaluators.

23. Here, '**closeness**' was measured using the **n-gram** algorithm. The human evaluators, on the other hand, evaluated the translation using a five-point scale of measurement. The automatic evaluation performed by **BLEU** was shown to be quite close to the evaluation performed by the **monolingual** human evaluators (Papineni et al. 2002).

24. A **test suite** consists of a carefully constructed set of examples that represent some pre-determined 'linguistic phenomena', meaning lexical and structural components.

25. A **test suite** can be used on a **single** linguistic phenomenon such as pronouns in an exhaustive and systematic way. A **test suite** is more useful for the evaluation of systems that have large syntactic and morphological analysis components. As a result, it is not always easy to construct an appropriate **test suite** that can test precisely what needs to be evaluated in a translation, where message and meaning are important. According to Prasad and Sarkar (2000), a **test suite** also has some weaknesses. While it has to be constructed manually to achieve systematic variation within a particular range of grammatical

phenomena, there is no standard method for constructing such a system. And since the same lexical items can be used repeatedly, findings can be misleading, resulting in an inaccurate evaluation (Arnold et al. 1994).

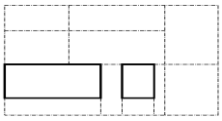
✍**26.** A **test corpus** is essentially a collection of texts which attempts to represent naturally occurring linguistic data. The **test corpus** methodology is based on the assumption that if a corpus is large enough, it is possible for any linguistic phenomenon of interest to occur at least once. Moreover, a **test corpus** can be used numerous times to test a variety of linguistic phenomena, and is usually also cheaper to construct than a test suite. Furthermore, a **corpus** can be compiled to reflect a user's needs.

✍**27.** Note that both these evaluation methods are **complementary** rather than competitive in nature as exemplified in the work of Prasad and Sarkar (2000).

✍**28.** The evaluation of natural-language processing tools can also be carried out using either a 'glass-box' or '**black-box**' approach.

✍**29.** A glass-box is sometimes referred to as a 'white-box', 'structural-box' or 'clear-box'. Its purpose is to test the **structural** components of a system by looking inside the box.

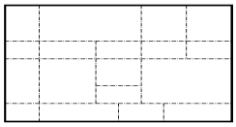
✍**30.** To illustrate this, let us imagine that a system is a rectangle, made up of smaller rectangles, as shown in *Figure 5.1*. These **smaller rectangles** are called **components** or **modules**. A glass-box evaluation is performed in order to **test specific components** of a system (rectangles highlighted in solid lines).



✍**31.** The approach is very useful to **researchers** and **developers** because they can identify the components that are experiencing problems. An example of a glass-box evaluation would be the testing of one or more components such as the parser, lexical look-up or semantic interpretation in a system (Palmer and Finin 1990: 177).

✍**32. Black-box** evaluation, on the other hand, is more suitable for **end-users**. It is also known as functional or behavioural testing. The evaluation focuses mainly on the overall performance of a system by looking at only the input (the source-language text) and output (the target language text), according to White (2003: 225). Imagine again that a system is a rectangle made up of smaller rectangles as shown in *Figure 5.2*. A **black-box** evaluation is carried out in order to test the

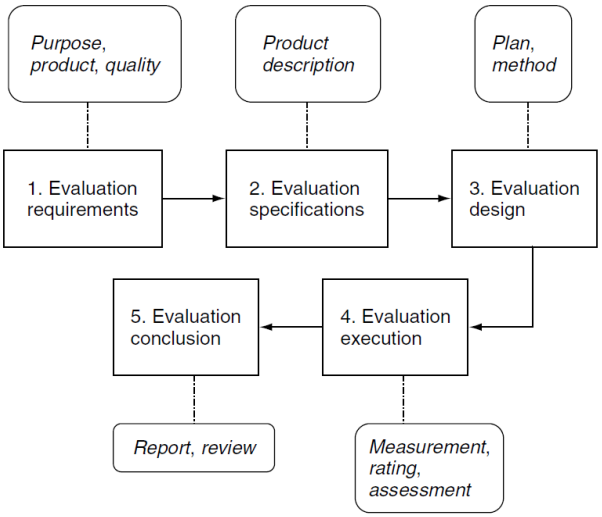
performance of the system as a whole (rectangle highlighted in solid lines).



33. The ISO 9126 series (Software Product Quality) provides definitions of six key **characteristics** used in evaluating the quality of software products:

- **Functionality**: meeting stated or implied needs of an end-user when functions of the system operate under specific conditions.
- **Reliability**: maintaining the level of performance by the system when operating under specific conditions.
- **Usability**: the ease of operating, understanding and learning each task of the system as a whole.
- **Efficiency**: the performance of the system in relation to the amount of resources available.
- **Maintainability**: the capability of the system to undergo modifications such as corrections, improvement and adaptations for different requirements and working environments.
- **Portability**: the ability to transfer the system from one environment to another such as to different operating systems.

34. *Figure 5.3* Example of an evaluation process:



35. FEMTI (Framework for the Evaluation of Machine Translation) evaluation framework:

❶ **Evaluation requirements**

(a) Evaluation purpose: to enable decisions to be made

- *Feasibility evaluation*: to discover if the approach used can be successful after further research and development. The stakeholders are mainly researchers and research sponsors.
- *Requirements elicitation*: to obtain reactions from potential stakeholders via the prototype system. The stakeholders are mainly developers and end-users.
- *Internal evaluation*: to perform periodic or continual evaluation at the research and development stage. The stakeholders are mainly developers and research sponsors.
- *Diagnostic evaluation*: to discover the causes of a system not producing the results as expected. The stakeholders are mainly developers.
- *Declarative evaluation*: to measure the ability of a system to handle a sample of real text especially the linguistic capability. The stakeholders are mainly researchers, developers and research sponsors.
- *Operational evaluation*: to discover if a system serves its intended purpose. The stakeholders are mainly researchers, developers and research sponsors.
- *Usability evaluation*: to measure the ability of a system to be useful to the intended end-user. The stakeholders are mainly developers and end-users.

(b) Evaluation objects: to identify the context of use, for example, a system that translates weather bulletins.

(c) Translation task characteristics: to discover from the endusers' point of view the purpose of the translation output. The stakeholders are mainly developers, research sponsors and end-users.

- *Assimilation*: to supervise the large volume of texts produced in more than one language. The stakeholders are mainly endusers.
- *Dissemination*: to deliver translations to others. The stakeholders are mainly end-users.
- *Communication*: to support speakers of different languages. The stakeholders are mainly end-users.

(d) User characteristics: to identify different groups of end-users.

- *Machine translation users*: interactions between end-users and the system. The stakeholders are mainly translators and post-editors.
- *Translation consumers*: end-users who use the translations.
- *Organizational user*: organizations who provide translations such as translation agencies.

(e) Input characteristics: to identify the format of the source language texts and information about the authors.

- *Text types*: genres, subject fields.

- *Authors*: level of proficiency in the source language and level of knowledge in the subject field.
- *Sources of errors*: linguistic errors and typographical errors.

❷ **System characteristics**

(a) Machine translation system-specific characteristics: components in the system and process flow of the system.

- *Translation process*: the underlying methodology behind the development of the system. In other words, how the knowledge of the translation process is represented and acquired in the system and when a particular type of knowledge is applied during the translation process. This also includes the language coverage such as dictionaries, glossaries, terminology databases and grammar.
- *Translation process flow*: the processes, such as pre-translation preparation, post-translation output and dictionary updating, that enable a system to operate successfully.

(b) External characteristics: the six qualities of translation based on ISO 9126 mentioned earlier in this chapter (functionality, reliability, usability, efficiency, maintainability and portability). In addition, cost is another characteristic that plays a major role in deciding whether a system can undergo a detailed evaluation.

Chapter 6

Recent Developments and Future Directions

✍ **1.** This chapter presents some recent developments and shows the direction in which translation technology is heading, in particular regarding the *future* of machine translation systems that are now incorporating speech technology features. The integration of **speech technology** and traditional machine translation systems allows translation not only between texts or between stretches of speech, but also between **text** and **speech**. This integration is proving to be useful in many specific situations around the globe especially in international relations and trade.

✍ **2.** This chapter also looks at research projects in countries that are involved in the development of translation tools for minority languages and discusses the problems encountered in developing machine translation systems for languages that are less well-known and not widely spoken. Another form of technology called the '**Semantic Web**' that has the potential to improve the performance of certain machine translation systems is also described. Included in this chapter, too, are issues such as **linguistic dominance** and **translation demands** on the WWW that are already shaping parts of the translation industry.

✍ **3.** The emergence of **data-driven methods** has motivated more research in the area of **natural-language processing**. Retrieving information from the **WWW** is currently achieved through the use of search engines such as **Google** (Macklovitch 2001).

✍ **4.** The question we now ask is, **what does the future hold for machine translation research?** One suggestion, from Schäler, Way and Carl (2003: 104), is illustrated in *Figure 6.1*:

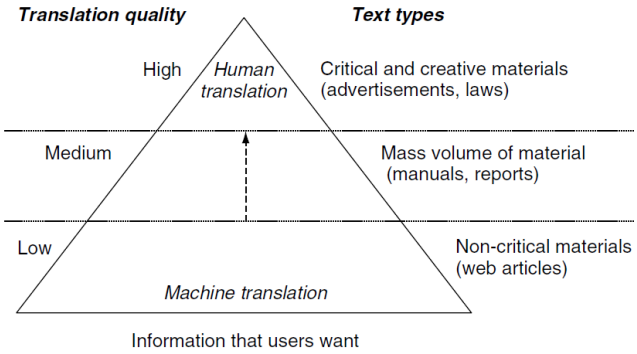


Figure 6.1 Future-use model of translation technology

✍ **5.** The model presented in the figure divides translation quality into three levels: **high**, **medium** and **low**, corresponding to three different types of text.

- ☆ At the **top** level, a human translator is chosen over a machine to translate texts that are of a creative nature where, for example, the use of metaphors is abundant. ***Examples*** of such texts include advertisements and plays.
- ◈ At the **middle** level, a combination of human translator and machine is used to translate large amounts of subject-specific texts for which accuracy and presentation, especially graphics, are important. ***Examples*** of such texts include laboratory reports and manuals.
- 🗎 At the **bottom** level, a machine is chosen over a human translator to produce rough translations at very little or no cost, such as those offered by online machine translation systems. ***Examples*** of non-critical texts are web pages for products, services and general information.

✍ **6.** As translation technology research and **development** progresses and produces more reliable systems, it is to be expected that the use of machine translation will increase as indicated by the **upward arrow**.

✍ **7.** In order to achieve **high-quality** translation, a machine translation system must be designed for a very **narrow subject** field where

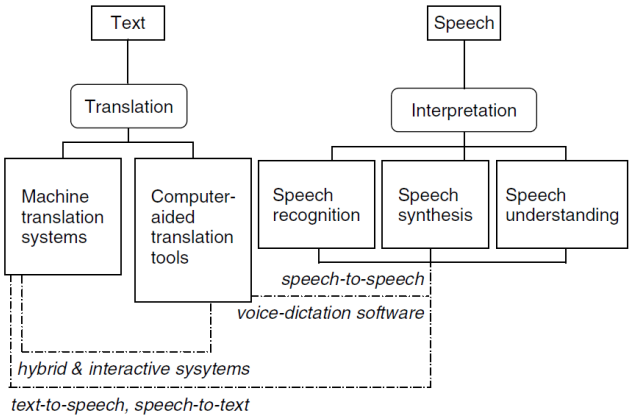
vocabulary and grammar is based on a **controlled language** (Sumita and Imamura 2002).

✍ **8.** Over a decade ago, the problems of complex sentence analysis, optimum target-language equivalents and generating idiomatic output had yet to be resolved, as pointed out by Hutchins (1994); this is still the case today. Furthermore, problems of **language analysis**, *transfer* and **synthesis**, *learning* and **common-sense reasoning**, especially for machine translation systems, are yet to be fully resolved.

✍ **9.** In the past, some professional translators orally recorded the first draft of a translation using what was called a '**dictaphone**' machine. From the dictaphone, the recorded translation draft was then typed up – usually by a secretarial assistant – to produce the written version of the translation. In contrast, the current technology of '**dictation equipment**', that is **voice-dictation software**, serves as an alternative to typing, having the ability to **convert recorded speech into text** automatically ('text' in this context always referring to written text). It can also be used to create, edit, revise and save translation documents based on the voice commands of the translator.

✍ **10. Voice-dictation** software is highly **language-specific**. Also, it is **voice-** and **accent-specific** necessitating training of the system by the *individual* translator.

✍ **11.** *Figure 6.2* shows the connection of speech technology to computer-aided translation and machine translation systems:



✍ **12.** A recent application of **speech-to-text** technology is in the conversion of spoken language on television into '**closed captions**' for the *deaf* and *hard-of-hearing* community, not strictly-speaking a translation application but one which has clear links to the human activity of interpreting.

✍ **13. Closed captions** are the written version not only of what is being said on television but also of relevant sounds such as 'PHONE RINGING' and 'FOOTSTEPS' for the benefits of people with a hearing disability. These captions can be activated by the viewer, sometimes with a special decoder.

✍ **14.** Examples of **speech-to-speech** translation systems include MASTOR (Multilingual Automatic Speech-to-Speech Translator) by IBM that is used to facilitate speech between individuals who share no common language, and HealthComm Healthcare Patient Communication Platform by Spoken Translation, Inc., which provides communication between Spanish-speaking patients and English-speaking healthcare workers.

✍ **15.** In recent years, several **less well-known** or '**low-density**' languages have also caught the attention of researchers working in the area of natural-language processing. In this field, **low-density** languages are those that have very few or no resources containing linguistic information, such as dictionaries and texts in an electronic format (McEnery, Baker and Burnard 2000). It means that any texts in print of a **low-density** language would have to be converted into an electronic format and be tagged before they can be of any use to the researchers in the area of machine translation development. Such an exercise is **time-consuming** and **costly**. Moreover, according to Jones and Rusk (2000), *commercial market forces* are unlikely to provide much incentive to work with **low-density** languages.

✍ **16.** One type of product that provides an instantaneous summarized translation is known as **automated real-time translation**. This system is especially useful to organizations where global communication is critical and time is of the essence in their daily business operations. Hence, instantaneous translations of **web pages**, documents, e-mails and other types of information are crucial. The system has several **advantages**, such as obtaining a translation of formal, as well as colloquial texts within seconds, rapid translation of foreign-language articles and real-time online communication (e-mails and chat-room messages) in multiple languages.

✍ **17.** The type of **English** language often used on the web has been described as a '*free-floating lingua franca*' or '**International English**', a language that has lost a large number of cultural and grammatical elements that tie it to its native speakers (Snell-Hornby 2000: 109).

✍ **18.** Two important points concerning developments in machine translation:

- that most machine translation systems are currently restricted in terms of subject fields and language pairs

- ❄ that the trend of future machine translation research seems to be moving towards **hybridization** between rule-based and corpus-based approaches.

✍ **19.** Current **online** machine translation systems generally produce **poor-quality** translations that do not reflect the real capabilities of the majority of machine translation systems. The reason is that almost all online machine translation systems rely on **limited** sets of linguistic **rules** of dictionary look-up and simple syntactic transfers following the **rule-based** approach. **Corpus-based** approaches are now seen as a serious challenge to the present rule-based online systems as a result of new technology. This new technology not only benefits current **online** machine translation systems but also current **corpus-based** and **knowledge-based** systems (Vertan 2004).

✍ **20.** As early as 1989, ***Tim Berners-Lee*** of the W3C, the creator of the WWW, HTML and other important web ideas, had already introduced the idea of what is now known as the '**Semantic Web**'. In 2001, after formalizing this idea, Berners-Lee together with his co-authors James Hendler and Ora Lassila defined the **Semantic Web** in *Scientific American* as 'an extension of the current Web in which information is given well-defined meaning, enabling computers and people to work in better cooperation'.

✍ **21.** With the **Semantic Web** – a universal medium for information exchange, providing meaning to the content of documents on the web that can be '**understood**' by machines – scientific communities, in particular the natural-language processing community, realized it had the potential to improve natural-language processing applications, especially machine translation systems.

✍ **22.** One potential beneficiary of the **Semantic Web** is example-based machine translation systems. Current **example-based** systems rely on generating new translations automatically via examples extracted from aligned parallel corpora. This approach is limited by the availability of such corpora, which are found only in certain subject fields and languages.

✍ **23.** Ever since **products** began to penetrate markets in different countries, the notion of translation and the nature of the translation industry have become increasingly complex. Products – including related documentation – that are to be sold in a specific market have to undergo certain changes as required by the trade regulations of that country. The **changes** involved concern not only the translation, for example of user manuals accompanying certain products, but also packaging; the **changes**

must be carried out in a manner **appropriate** to the target market, a process known as '**localization**'.

✍**24.** It is important to remind ourselves that **no** technology can entirely replace **human translators**, for the simple reason that **humans** are still needed to produce high-quality translations. Human languages are **multilayered** in *usage* and *meanings*, and current technology remains unable to decipher the finer nuances of human languages in the same way as humans can. Technology is restricted to its specific uses and, as a result, is destined to remain as a tool.

Chapter 7
Translation Types Revisited

✍**1.** The book concludes by presenting an expanded version of the *four* basic **classifications of translation types** as suggested by Hutchins and Somers (1992) and introduced in Chapter 1. It is concluded that the one-dimensional linear continuum originally proposed is no longer able to accurately reflect current developments in translation technology.

✍**2.** Translation tools today come in different versions and types depending on the purposes for which they are built. Some are multifunctional while others remain **monofunctional**. An alternative way must therefore be found to depict the complexities and multidimensional relationships between the **four translation types** and the topics discussed in this book.

✍**3.** It is not possible to put every single subject discussed here into one diagram or figure, and so, in order to gain a better understanding of how the issues are related to one another, they are divided into groups. Topics or issues in each group have a common theme that links them together, and are presented in a series of tables. However, it is important to bear in mind that not all topics can be presented neatly and easily even in this way. This clearly shows the complexity and multidimensionality of **translation activities** in the modern technological world.

✍**4.** At the end of the book, several Appendices provide information on the various **Internet sites** for many different translation tools and translation support tools such as monolingual, bilingual, trilingual and multilingual dictionaries, glossaries, thesauri and encyclopaedia.

✍**5. Machine translation** (MT) systems are purely automatic with no human intervention during the actual translation process. They are conventionally divided into specific-purpose systems for highly specialized technical and subject-field-specific texts on the one hand, and general-purpose systems for general-purpose texts on the other hand. The

general-purpose systems now also include online machine translation systems found on the Internet.

6. Human-aided machine translation systems (HAMT) are essentially a form of machine translation with an interactive mode; the principal contribution to the translation is made by the machine but a human can intervene during the translation process.

7. Computer-aided translation (CAT) includes translation tools, linguistic tools and localization tools such as translation memory systems, electronic dictionaries and concordancers; the translator makes a much greater contribution here than in HAMT.

8. Human translation (HT) refers specifically to translations performed by translators.

9. *Table 7.1* **Degree of automation**:

	MT		HAMT	CAT	HT
	Specific	General			
Fully automated	Y	Y	N	N	N
Partially automated	N	N	Y	Y	Y
Non-automated	N	N	N	N	Y

Y = yes; N = no

10. When a system has only partial automation, the element that completes it will be the human element. Clearly, '**intervention**' is not applicable to human translation.

11. *Table 7.2* **Human intervention**:

	MT		HAMT	CAT	HT
	Specific	General			
Human intervention	N	N	Y	Y	n/a

Y = yes; N = no; n/a = not applicable

12. *Table 7.3* shows which tools and technologies can be integrated with other tools/systems.

Table 7.3 Integrated tools

	MT		HAMT	CAT	HT
	Specific	General			
Machine translation	n/a	n/a	Y	Y	n/a
Translation tools	Y	Y	Y	Y	n/a
Linguistic tools	Y	Y	Y	Y	n/a
Localization tools	P	P	P	Y	n/a
Speech technology	Y	Y	Y	Y	n/a

Y = yes; P = possible; n/a = not applicable

✍ **13.** *Table 7.4* Application of theory:

	MT		HAMT	CAT	HT
	Specific	General			
Translation theory	N	N	N	N	P
Linguistic theory	Y	Y	Y	Y	P

Y = yes; N = no; P = possible

✍ **14.** *Table 7.5* Application of theory in machine translation systems (The description in *Table 7.5* applies only to traditional machine translation and not to hybrid and integrated systems):

	Direct translation	Rule-based		Corpus-based	
		Interlingua	Transfer	Statistical	Example
Translation theory	N	N	N	N	N
Formal linguistic theory	M	Y	Y	N	Y

Y = yes; N = no; M = minimal

✍ **15.** A **general-purpose** machine translation system is likely to perform less well than a **specific-purpose** system even with a controlled language source text, simply because a general-purpose system is not designed to translate texts from narrow subject fields. **General-purpose** systems can cope with a broader range of input texts but with expectations of lower quality.

✍ **16.** The **quality** of a ***target-language text*** produced especially by semi- or automated systems mostly hinges on a number of factors such as

- 🗐 the coverage provided by the **dictionary** or dictionaries in a system
- ❋ the coverage of terms in a **terminology database**
- ▣ the capabilities of the **analysis** and **synthesis** modules in a system
- ✺ the **quality** of the *source-language text*, as well as its **type**

17. *Table 7.7* shows for which translation type the **target-language text** may need to undergo **post-editing** to produce the required quality of translation.

- **Rapid post-editing** can be performed on target-language texts generated by both specific- and general-purpose machine translation systems, and also human-aided machine translation systems where the text is needed for information only according to a specific purpose, for a specific group or for a specific period of time.
- **Polished post-editing**, on the other hand, is almost always required for the translations generated by specific-purpose machine translation systems: where the subject matter is highly technical, such as in operational manuals, accuracy and clarity are crucial.

Table 7.7 Target-language texts

	MT		HAMT	CAT	HT
	Specific	General			
Post-editing:					
Rapid	R	R	R	n/a	n/a
Polished	R	MR	R	n/a	n/a
Revision	n/a	n/a	n/a	R	R

R = required; MR = may be required; n/a = not applicable

18. In *Table 7.8* we show which task, performed at a certain stage of the translation process, is important to which translation type. We extend the meaning of '**interactive**' here beyond the conventional understanding of human–machine interaction in human-assisted machine translation to the use of any tool involving both human and machine.

	MT		HAMT	CAT	HT
	Specific	General			
Pre-editing	I	LI	I	LI	LI
Interactive	n/a	n/a	I	I	I
Post-editing	I	LI	I	n/a	n/a

I = important; LI = less important; n/a = not applicable

19. In *Table 7.9*, some different types are examined, whereby '**type**' is described on a scale from *highly* **creative** to *highly* **technical**:

	MT		HAMT	CAT	HT
	Specific	General			
Highly creative	NS	NS	NS	NS	S
Semi-creative	NS	NS	NS	NS	S
General-purpose	NS	S	P	S	S
Semi-technical	S	P	S	S	S
Highly technical	S	NS	S	S	S

S = suitable; NS = not suitable; P = possible

20. Highly creative persuasive texts such as advertisements or expressive texts such as poems are **not suitable** for either specific- or general-purpose machine translation systems for a number of reasons, including novel or unconventional uses of language such as non-standard syntax patterns or neologisms, for which there is no equivalent word or phrase in the other language. In contrast, **semi-technical** and **highly technical** texts are the **most suitable** types of text for specific-purpose machine translation systems.

21. Some tools are designed for ***specific*** languages. ***Spell-checkers*** are an obvious example, as are also electronic dictionaries and glossaries. Others, such as **translation memory** systems and **concordancers**, can be used with **any language**, assuming that the relevant character sets are digitally available.

22. *Table 7.10* reviews each translation type with respect to their degree of **independence** from particular languages:

	MT		HAMT	CAT	HT
	Specific	General			
Language-dependency	H	H	H	H/L	H

H = high; L = low

23. Our last perspective on **language-dependency** – see Table 7.11 – concerns **controlled language**, for example for highly specific purposes such as ASD Simplified English, compared with 'natural' language, as in standard British English for example.

Table 7.11 Types of source language

	MT		HAMT	CAT	HT
	Specific	General			
Controlled language	Y	P	Y	Y	P
Natural language	N	Y	P	Y	Y

Y = yes; N = no; P = possible

24. For **specific-purpose** machine translation systems, **controlled language** is more suitable for source-language texts than is **natural language**. On the other hand, **natural language** is best suited for general-purpose machine translation systems. **Controlled language** can facilitate both human-aided machine translation and computer-aided translation. The same, however, cannot be said about **natural language**, which is more suitable for computer-aided translation than for human-aided machine translation. Both varieties of language are acceptable to human translators. For stylistic and other reasons, a **natural language** text presents more of a challenge than a restricted **controlled language** text and human translators may therefore prefer it.

25. *Table 7.12* shows how important standards are for each translation type, referring to three different types of standard: TMX, TBX and XLIFF.

Table 7.12 Data interchange standards in translation

	MT		HAMT	CAT	HT
	Specific	General			
Translation standard (TMX)	P	P	I	VI	VI
Terminological standard (TBX)	VI	VI	VI	VI	VI
Localization standard (XLIFF)	P	P	I	I	I

VI = very important; I = important; P = possible

26. *Table 7.13* Translation groups and data interchange **standards**:

	Translation standard (TMX)	Terminological standard (TBX)	Localization standard (XLIFF)
Professional translators	VI	VI	P
Translation companies	VI	VI	VI
Localization industry	VI	VI	VI
Researchers	I	I	I
Developers	I	I	I

VI = very important; I = important; P = possible

27. '**Evaluation**' is a term applied to the assessment of translation output from automated systems; so evaluation in this sense is **not applicable** to the work of human translators.

28. *Table 7.14* Levels of **evaluation**:

	MT		HAMT	CAT	HT
	Specific	General			
Components in system	I	I	I	I	n/a
Whole system	I	I	I	I	n/a

I = important; n/a = not applicable

29. *Table 7.15* **Methods of evaluation**:

	Components in a system	Whole system
Human	Y	Y
Automation	Y	Y
Test suite	Y	N
Test corpus	Y	N
Glass-box	Y	N
Black-box	N	Y

Y = yes; N = no

30. The aim in *Table 7.15* is to provide a clearer perspective on which evaluation methods are better suited to test an individual component in a system or an entire system. Each method uses different variables or test material to perform the evaluation. Some test material consists of *linguistic phenomena* that have been artificially created to evaluate particular features of the system (**test suite**), while other test material is extracted from a corpus (**test corpus**). A variety of **tests** is used to evaluate a specific component of a tool during its developmental stage, whereas to evaluate an entire system, the most suitable methods are human judgement, automation and black-box.

31. *Tables 7.16* and *7.17* show the different approaches to machine translation such as direct translation, **ruled-based** and **corpus-based** approaches as they relate to particular design features and coverage of language pairs respectively. Rule-based and corpus-based approaches are further divided into their respective sub-types.

32. *Table 7.16* Features in a machine translation system:

	Direct translation	Rule-based		Corpus-based	
		Interlingua	Transfer	Statistical-based	Example-based
Algorithms	N	N	N	Y	Y
Examples	N	N	N	N	Y
Dictionaries	Y	Y	Y	N	N
SL analysis	M	Y	Y	N	N
TL synthesis	M	Y	Y	N	N
Abstract representations	N	Y	Y	N	N
Transfer module	N	N	Y	N	N
Language model	N	N	N	Y	Y
Translation model	N	N	N	Y	N
Modularity	N	Y	Y	Y	Y
Corpora	N	N	N	Y	Y

Y = yes; N = no; M = minimal

33. One feature found in nearly all machine translation systems from the second generation onwards (**rule-based** and **corpus-based** systems) is **modularity**, meaning that the components of the system are independent of each other so that a researcher can change or improve a particular module without this affecting the performance of other modules of a system.

34. Modularity is desirable in a machine translation system as it can reduce development and maintenance costs when new language pairs are added (*Table 7.17*). A feature that is also important to machine translation development is the reversibility property that enables a language pair working in one direction to be reversed.

	Direct translation	Rule-based		Corpus-based	
		Interlingua	Transfer	Statistical	Example
One language pair	Y	Y	Y	Y	Y
More than one language pair	N	Y	Y	Y	Y

Y = yes; N = no;

35. Unlike other translation tools, **all** machine translation systems are **language-dependent** and contain minimally one language pair.

36. *Table 7.18* **Texts** and computer-aided translation **tools**:

	Translation tools	Localization tools	Linguistic tools
Highly creative	U	NU	VU
Semi-creative	U	NU	VU
General	U	NU	VU
Semi-technical	VU	VU	VU
Highly technical	VU	VU	VU

VU = very useful; U = useful; NU = not useful

37. Localization tools are **not** designed for the translation of semi-creative, highly creative and general-purpose texts. They have been developed to deal with **technical texts** such as product specifications and instruction manuals. Whether translation or linguistic tools are useful in the translation of general-purpose texts may depend on the translator, the degree of ambiguity and the purpose of the translation.

38. Unlike machine translation systems, **translation** and **localization tools** are *rarely* language-dependent, whereas some linguistic tools such as spell-checkers, grammar checkers and dictionaries can be language-dependent. Concordancers, on the other hand, tend to be language-independent, as shown in *Table 7.19*.

Table 7.19 Language dependency in computer-aided translation tools

	Translation tools	Localization tools	Linguistic tools
Language-dependency	L	L	H/L

H = high; L = low

Short Answer Items & Tests

1.2 Short Answer Items

1. After the ……. translation approach, which had much in common with a word-for-word approach to translation owing to the central role of dictionaries in the system, the next generation of systems–known as '…….' systems–make use of a number of formal grammars in the design of machine translation systems.

2. The type of machine translation labeled '…….' is designed specifically for the translation of electronic documents obtained from the Web.

3. A generally accepted view of ……. translation is 'a system wherein the computer is responsible for producing the translation per se, but may interact with a human monitor at many stages along the way' (Slocum 1988).

4. Integrated machine-aided human translation systems are sometimes known as '…….', as they combine a number of tools.

5. In the early days of translation theory, Nida's idea of the translation process as working from the source text to the target text by reaching down to an underlying level of meaning as the means of '…….' between the languages resonates with Chomsky's model.

6. Skopos theory allows a source-language text to be translated into a number of different target-language texts depending on the ……. specified in the so-called 'translation …….' or brief.

7. ……. translation has been described as the use of computer software by translators 'to perform part of the process of translation' (Sager 1994).

8. Common ground between Translation Studies and translation technology–and machine translation in particular–may be found within ……. approaches to translation.

9. Until the late 1960s, the method used to generate translations in nearly all machine translation systems was the '……. translation' approach. This approach is based on the assumption that one target-language word can be generated from one source-language word. It also requires a ……. analysis, for example, recognition of word classes such as noun and verb (Hutchins 1979).

10. According to Trujillo (1999), the ……. theory of translation strategy, for example, 'arose as a response to the growing need for non-

literary translation'. The focus on the purpose of the target text in relation to its translation setting resonates with a common definition of translation quality as 'fitness for …….'.

1.3 Answers

1) direct, rule-based	**2)** Online
3) human-aided machine	**4)** workstations
5) transfer	**6)** purpose, commission
7) Machine-aided human	**8)** functional
9) direct, minimal syntactic	**10)** Skopos, purpose

☙☆❧ 1.4 Tests ☙☆❧

✍ Select the best choice.

1. The size of the dictionaries and the capabilities of the syntactic ……. **components generally indicate how good a system is.**

a) analysis (not synthesis) b) synthesis (not analysis)
c) analysis and synthesis d) non of the above

2. The type of machine translation labeled '…….**' refers to machine translation systems for home users who have few or no translation skills.**

a) Naïve b) Apprentice
c) Home d) Novice

3. A ……. **text is one that has been edited by a human, in most cases by someone other than the author, prior to the translation process, whereas a** ……. **text is usually written following certain strict linguistic rules.**

a) pre-edited, controlled-language b) post-edited, controlled-language
c) controlled-language, pre-edited d) post-edited, pre-edited

4. ……. **machine-aided human translation systems are sometimes known as '**…….**', as they combine a number of tools.**

a) Integrated, workbenches b) Non-integrated, workstations
c) Non-integrative, workbenches d) Integer, workbenches

5. In MT, ……. **refers to the process of changing the documentation of a product, a product itself or the delivery of services so that they are appropriate and acceptable to the target society and culture.**

a) multitasking b) integration
c) parsing d) localization

6. The period between the 1950s and the 1960s saw the return of the dichotomy of oppositions similar to that of word-for-word versus sense-for-sense such as 'formal versus …….**' as proposed by Eugene Nida (1964), where the** ……. **leans toward the source-language text structures while the** ……. **adapts the translation more closely to the target language in order to achieve** ……. **.**

a) dynamic, latter, former, naturalness
b) functional, former, latter, accuracy and fidelity
c) dynamic, former, latter, naturalness
d) functional, latter, former, accuracy and fidelity

7. In the late 1970s, a dichotomy was introduced by Juliane House in the form of "…….". While in "……." translation, it is clear that the target-language text is a translation from another language, "……." translation does not show that the target text originates in another language.

a) overt versus covert, covert, overt
b) overt versus covert, overt, covert
c) formal versus dynamic, dynamic, formal
d) formal versus dynamic, formal, dynamic

8. Vermeer's Skopos theory draws heavily on the 'translational ……. **theory' developed by Justa Holz-Mänttäri, which represents a** …….**-oriented approach to the theory and practice of translation.**

a) action, function
b) relevance, function
c) operation, target
d) operation, source

9. The emergence of ……. **theory is seen as part of a general shift from predominantly linguistic based translation theories to a theory that has an orientation towards the way a translation functions in the target society and culture.**

a) Relevance
b) Polysystem
c) Skopos
d) Manipulation

10. In the early 1980s, Peter Newmark introduced the dichotomy of '……. **translation', which follows as closely as possible the semantic and syntactic structures of the source language text, and '**……. **translation', which is focused on the reader and 'attempts to produce … an effect as close as possible to that obtained on the readers of the original' (Newmark 1981), recalling Nida's well-known '**……. **equivalence'.**

a) communicative, semantic, dynamic
b) semantic, communicative, formal
c) communicative, semantic, formal
d) semantic, communicative, dynamic

☙✷❧ 1.5 Answer key ☙✷❧

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

Book ❷

Translation-mediated Communication in a Digital World:

Facing the Challenges of Globalization and Localization

M. O'Hagan & D. Ashworth

2.1 Notes

Introduction

1. Our main hypothesis is that **technological changes** affecting communication modes are going to profoundly impact on the *professions* of **translators** and **interpreters** to such an extent that new professions will result. Our assumption is that new modes of communication employed across languages will both drive and enable new types of language support.

PART I

Setting the Scene

Part I provides the big picture, highlighting the major changes taking place in translation and interpretation (which we refer to as Translation to include both) with the advent of the Internet. We introduce a new framework Translation-mediated Communication (TMC).

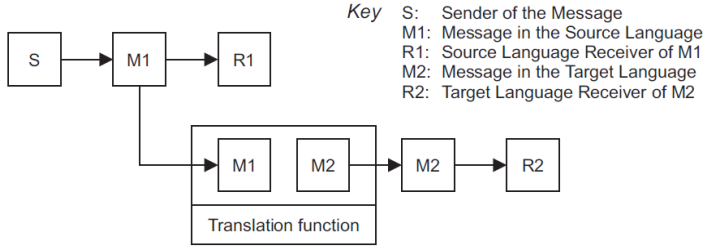
Chapter 1

Translation and Interpretation in Transition: Serving the Digital World

1. ***Chapter 1*** describes the traditional function of Translation on the basis of TMC, and highlights issues arising from a newly emerging context in which Translation has to function.

2. We use TMC as the framework for our exploration and this in turn means that we take the approach of treating **Translation as communication**. There are a number of scholars who developed

translation models based on *Shannon's* ***Mathematical Model of Communication***, including Nida & Taber (1969), Bell (1991) and Gile (1995). The simplicity of the Shannon model allows us to illustrate the role of Translation as an embedded function between the **sender** and the **receiver** with the Translator acting both as the **receiver** of the message in the source language and the **sender** of the message in the target language as described by Nida and Taber (*Figure 1.1*).



✍**3.** This model highlights the **purpose** of Translation as 'an act of communication which attempts to relay, across cultural and linguistic boundaries, another act of communication…' (Hatim & Mason, 1997:1).

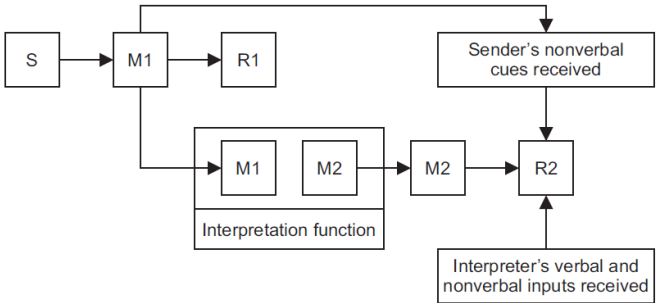
✍**4.** Given that ***Shannon's*** model was originally intended for synchronous telephone communication, this model is equally applicable to **interpretation**, in which the sender and the receiver may be engaged in constant turn taking.

✍**5.** The main difference in *modus operandi* between **translation** and **interpretation** resides in the fact that interpretation caters to synchronous communication where all communicating parties (including the interpreter) are normally present in one physical location and communicate in real-time. By comparison, translation facilitates asynchronous communication via writing with a certain time lag.

✍**6.** Given that ***Shannon's*** Communication model tends to focus on the transmission function of telecommunications, we will combine our analysis of TMC with ***Gile*** (1995), who also uses a communication-based approach but is more focused on the **Sender**, the **Receiver** and the **Message**.

✍**7.** *Figure 1.2* shows the interactions among the **Sender**, the **Receiver** and the interpreter in a typical small group *face-to-face* **consecutive interpreting** situation. It illustrates how the Receiver (R_2) observes the Message (M_1) being delivered by the Sender (S), albeit without understanding the verbal content but taking in some nonverbal communication cues such as facial expressions and body movements (kinesics), although the Receiver (R_2)may not 'read' them correctly. This

contrasts with the situation for the translator, who normally works in isolation from either the Sender or the Receiver.



8. In conventional Translation, the Message consists of written texts for translation and speech for interpretation. Gile (1995: 26) sees it consisting of '**content**' and '**package**'. The term '**package**' refers to 'the linguistic and peri-linguistic choices made by the **Sender** and to the physical medium through which they are instantiated.'

9. According to Gile's definition, in written texts, the **package** will include words, grammatical structures, fonts, page layout, graphics, etc. For speech, it is made up of the words, grammatical structures, the voice and delivery, as well as nonverbal cues.

10. **Content** and **package** interact to affect the **message** as a whole. As pointed out by Gile, a good content can be weakened by poor style of writing or delivery of speech, and vice versa. In thinking of the change in the nature of the **Message** with the advent of the Internet, this dual view to analyze the **Message** becomes relevant to our purposes to highlight the changing nature of the **Message**.

11. As pointed out by Gile (1995: 32), 'the Translator is instrumental in helping to achieve the Sender's aims, but cannot guarantee their fulfillment.' This may be evident if one considers communication break-downs that commonly take place between sender and receiver speaking the same language and sharing the same cultural background. In other words, the successful facilitation of **inter-lingual communication** is not entirely determined by the performance of the Translator alone, but also is affected by the **Sender**, the **Message** and the **Receiver**.

12. One significant source of failure in **inter-lingual communication** can stem from incorrect **assumptions** of common beliefs and experience that actually differ according to cultural background, knowledge, preferences and pragmatics (use of language). A *US* learner of *Japanese* may interpret a negative question 'Aren't you going to the theater?' as asking

for a yes-or-no answer (seeking facts) when it is an invitation (in context).

✍ **13.** Translation work may be commissioned by ❶ the **Sender** or ❷ the **Receiver** of the Message:

❶ When the **Sender** of the Message commissions the translation work, the Message reaches the Receiver in his or her language. **Localization** is a good example of this, as it attempts to adapt the Message to be suitable to the Receiver of the message. Literary translation is another example of **Sender-commissioned** translation – it enables the Receiver to read literary works in his or her own language. This pattern tends to take place for information dissemination rather than gathering purposes and a high-quality translation is generally required. However, in this case, the **Sender** of the Message is typically unable to directly assess the quality of the translation, thus the feedback on translation tends to come from the Receiver (end-user) to the translator normally via the Sender.

❷ When the **Receiver** commissions the work, it is because the message is received in an unfamiliar language. For example, a Japanese scientist who receives the abstract of a technical paper in German may decide to have it translated into Japanese. In this case, the translation exercise normally does not directly affect the Sender of the Message. This pattern tends to take place for information-gathering purposes, and may not always require a top-quality translation. Recent applications of machine translation (MT) to browse Web pages in real-time are an example of this – the **Receiver** of the Message needs a translation, and the quality required is often for 'information only' purposes. In this case, the **Receiver** is able to provide feedback directly to the provider of the service or the translator, as the target language is the Receiver's language.

✍ **14.** The following *features* may characterize the **text** used in a **Web site**:

✵ the readership of the text is unspecified and can mean an extremely wide range of native speaker population;

◈ the text will be read on screen rather than on paper, at least in the first instance;

⌘ the text may be read in any order, and therefore in different contexts, depending on which hyperlink the reader may follow;

✪ the text is subject to much more frequent changes than is paper-based text;

※ the text may need to be 'adapted' to the target market readers, involving content changes; and

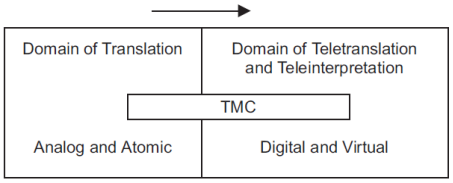
☆ the text may contain multimedia components, such as audio and extensive graphics and icons, whose cultural appropriateness may need to be considered against the target-culture norms.

Chapter 2

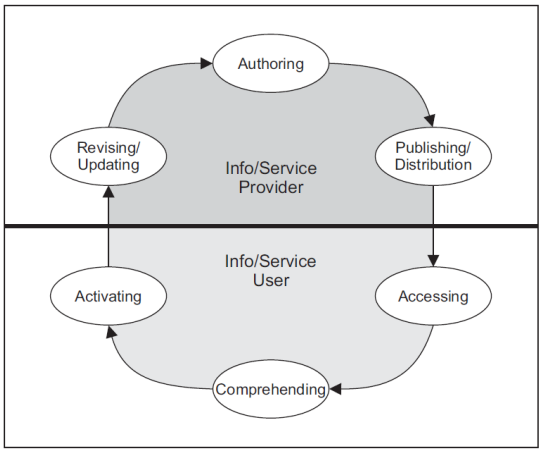
Redefining Context for Teletranslation and Teleinterpretation

✍**1.** ***Chapter 2*** concentrates on a number of specific attributes of the emerging context in the shift to **teletranslation** and **teleinterpretation**. The new context includes key concepts such as **digital literacy**, particularly in light of translation and translator *competence*. Changes are also considered by describing Translation as a communication system.

✍**2.** We may illustrate the redefinition of the **translator's workspace** as shown in *Figure 2.1*, where the framework based on TMC becomes more applicable as we move to teletranslation and teleinterpretation:



✍**3.** *Figure 2.2* illustrates how **digital content**, such as Web documents, goes through its lifecycle. This is what we call the 'digital content lifecycle' based on the InfoCycle by Lockwood (1998). *Figure 2.2* shows key nodes involved in the *lifecycle* of **digital content**. It starts from authoring of text, which may include multimedia elements, followed by distribution. The user of such an information service accesses it via some kind of IT device and understands the **content** before taking some sort of action based upon the information. The information provider will use *feedback* from the customers to revise and update the content. Within this cycle, language support may be required at almost any point. For example, **localization** is typically applied after authoring and before distribution, whereas WebMT may be used to translate the specified site on the fly during the comprehension stage:



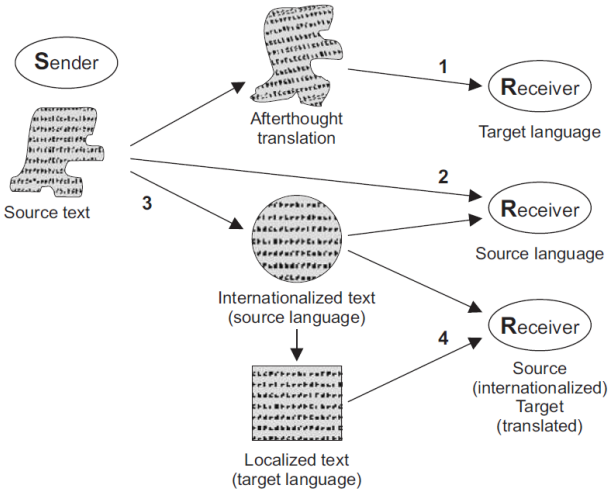
4. Sager (1993: 211) stresses the role of **Translation**: 'as a commissioned task, which starts with a need for communication and ends with a finished product.'

5. It is useful to think about the requirements of Translation in terms of '**translation competence**' and '**translator competence**' as distinguished by Kiraly (2000a).

- The *former* refers primarily to the competence to produce acceptable translations, however one might define 'acceptable.'
- The *latter* term refers to the skills and knowledge a translator needs in addition to translation competence. One can argue that the nature of electronic documents has an influence on both translation competence and translator competence.

6. **Teletranslators** and **teleinterpreters** need a wide range of *knowledge* and *skills* to be literate in the **digital environment**. In addition to translating or interpreting the conventional Message in the conventional mode, they will increasingly be involved in translating such digital contents as *software*, *web pages*, and *multimedia*.

7. *Figure 2.3* **Internationalization process**:

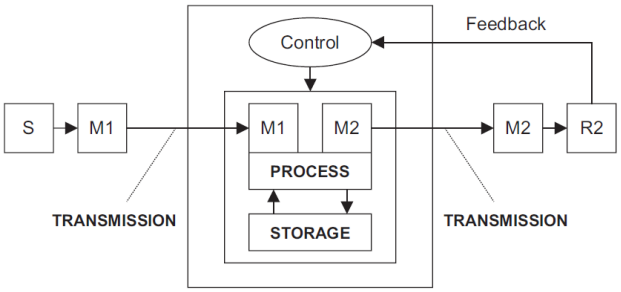


8. In 1 'afterthought translation' means that the source language text is not prepared with translation in mind. This may result in awkward translation for the target language Receiver in large parts resembling the source language (translatese). **In 2** the Receiver is a speaker of the source language, so there is no translation required. **In 3** the source language is modified by being internationalized, and is understandable both to a Receiver who speaks the source language and to a Receiver for whom it is a second language. **In 4** the internationalized text is further localized to become intelligible to other Receivers, who do not know the source language.

9. The **internationalization process** makes the Message more amenable to the subsequent translation into the Receiver's language. Note that the form of the Message changes, as represented by different shapes. The different shapes illustrate certain preferences by the culture of the target language. In this particular case, the culture of the target language prefers squares to circles.

10. Dealing with digital content means that the Message is provided in a **machine-readable form**, which automatically facilitates the use of certain tools. MT is probably the best example, although it is hardly used as a regular '**tool**' by most translation providers. Other tools include Translation Memory, terminology management systems, electronic dictionaries, and online databases, all of which work most efficiently with digital text. Otherwise the text needs to be ***retyped*** or to go through ***OCR*** (Optical Character Recognition), which is time-consuming and may introduce errors. Another important tool is management software of various kinds, such as ***workflow*** programs, including ***translation manager*** programs.

11. *Figure 2.4* shows a Translation Communication System with the underlying communication flow between the Sender (S) and the Receiver (R_2) of the Message (M_2) in the target language. The **Translator** carries out their function by applying their **internal knowledge** stored in their biological memories as well as in *auxiliary memories* such as dictionaries and databases in order to process (translate or interpret) the given **Message** (written or spoken), which is transmitted to the **Receiver**. For such a system to function, it also needs a ***control*** *function*, which takes in *feedback* and maintains the system as a whole.



PART II
Technologies Enabling Teletranslation

Part II concentrates on technologies, which are both driving and enabling new forms of Translation, together with wider implications of globalization and localization.

Chapter 3
Language Engineering and the Internet

1. ***Chapter 3*** looks into natural language **processing** technologies that have become particularly relevant to the digital communications environment on the **Internet**. It discusses the developments of language support automation, and tools for translators and interpreters.

2. The term '**language engineering**' is explained by the European Commission as follows: 'language engineering applies knowledge of language to the development of computer systems that can recognize, understand, interpret, and generate human language in all its forms' ('The Doctor is in,' 1998).

3. Many translators have purchased **Translation memory** (TM) systems at considerable expense in order to make it easier to carry out large projects. Since **TM** is based on collections of parallel texts, which

compare and match a source and target translation, it has the potential for creating very large, **context-sensitive** databases that can be of immense value to any translator, as long as these databases can be accessible. One of the primary **problems** confronting any translator is the use of language in **context**. For this reason, bilingual glossaries and dictionaries are usually very inadequate, especially if they contain only single-word definitions.

4. Extraction tools allow the translator to view a Web document without having to see the underlying HTML, XML or JavaScript code. Since the translators are able to see only the texts, they will not be confused by having to search through the code to find the source text that needs to be translated. **Extraction tools** are not normally provided as independent programs for downloading or purchase, but reside on the server of the translation service provider and must be used for the particular client. Some recent versions of *translation memory* products have also incorporated such tools.

Chapter 4
Computer-mediated Communication and Translation

1. ***Chapter 4*** turns to technological developments that are driving an underlying change in communications modes, notably Computer-mediated Communication (**CMC**) modes. On the basis of specific characteristics of the **CMC** mode, we introduce a potential *hybrid* language support called **transterpreting**.

2. One of the significant impacts of the **Internet** on **Translation** is the changing nature of the **Message** and the way in which it is transmitted, stored and processed.

3. Flanagan (1997) categorizes **online texts** into three general groups of

- ‘reference text’
- ‘communicative text’
- ‘interactive text’

4. Web documents are used for reference and for information dissemination and gathering purposes, and normally form uni-directional communication where the Receiver accesses the site to view already-existing materials. **E-mail messages** are sent to individuals, forum groups or newsgroups for communicative purposes because they elicit responses. **Chat** takes place in real-time, mainly via typed text between two or more individuals, although the voice channel is being incorporated in some chat platforms. These **CMC** modes can also be

classified in terms of **asynchronous** (e.g. Web and e-mail) and **synchronous** (e.g. chat) CMC.

✍**5.** *Table 4.1* **CMC** according to online text types

Online text category	*Reference text*	*Communicative text*	*Interactive text*
Online text examples	Web documents	e-mail messages	chat messages
Mode of communication	asynchronous		synchronous
Multimedia components used for presenting information	hypertext	linear text with embedded hyperlink	linear text
	2D and desktop 3D graphics, video	emoticons and other ASCII-based images	
		multimedia as attached file, including voice	2D and desktop 3D computer graphics
	audio		voice

✍**6.** The **Web** is constructed on the basis of hypertext, in which related information is tagged. The information is stored in different physical locations (servers), but the link is made seamlessly from the user's computer (client) via **HTTP** (Hypertext Transfer Protocol) using the point-and-click mechanism. Web texts cover a wide range of topics.

✍**7. E-mail messages** may be characterized in terms of their similarity to spoken **communication**, which is likely to contain sentence fragments, misspellings, misused punctuation and online jargon (Herring, 1996). *Table 4.2* shows distinctively Japanese emoticons that are closely tied to physical nonverbal cues used in Japanese **communication**. The uniquely Japanese '***cold sweat***' emoticon is used in contexts in which the writer of the message is concerned that the message in question may offend the recipient. This directly reflects the Japanese style of communication, which values the sign of modesty that indicates that the writer of the message fears that the message may be too opinionated.

Table 4.2 Examples of Japanese-specific emoticons

(^·^)	a smile with a dot for a mouth, since it is impolite for women to show their teeth
\(^_^)/	a 'banzai' [cheers] smile with arms upraised in a Japanese gesture
(^^;)	a cold sweat – among the most commonly used emoticons in Japan
(_o_)	a person sitting with head down and hands stretched in front in a typical gesture for a deep apology

✍**8. Chat messages** can be characterized by features such as:

- addressivity (e.g. including the name or abbreviation of the addressee in one's message),

- abbreviations
- paralinguistic and prosodic cues and
- actions and gestures (Werry, 1996).

9. Examples of **abbreviation** may include IMHO (in my humble opinion) or TTYL (talk to you later). The reason for their use is attributed to 'screen size, average typing speed, minimal response times, competition for attention, channel population and the pace of channel conversations' (Werry, 1996: 53). **Abbreviation** makes for *compactness* and *brevity* and is therefore easier to type, in turn achieving a higher response speed. This suggests that the **translator** serving this mode of communication requires knowledge of such abbreviations in both the source and target languages.

10. An early example of extending an **interpreting** service to the text-based chat environment was demonstrated in the Community Access 96 conference held in November 1996 in Nova Scotia, Canada in which computer conferencing was used to connect remote participants via IRC (Ashworth, 1997). The conference was conducted in Canada's two official languages, English and French. Registered participants who did not attend the conference in person could view the transcript of the speeches in both languages. In addition, they could discuss the topics among themselves via chat, and submit questions to the attendees. These were **interpreted** by having an interpreter stand behind a typist, who would type in the translation provided by the interpreter.

11. About one month prior to this development, ***Ashworth*** (1997) conducted a pilot experiment, which he called '**transterpreting**' in which, unlike the above example, the translated text was input directly by the '**transterpreter**.' The **transterpreter** used two terminals, each of which showed the chat dialogue in a single language. This was to overcome the character encoding problems that make it difficult to display single-byte, and double-byte characters side by side (unless one could use a *Unicode-based* platform, which was not available). Each chat participant therefore saw only the translated chat line from the respective partner. In this study, Ashworth realized the difficult problems involved in trying to provide **simultaneous transterpreting**, particularly between English and Chinese. Japanese was also a problem, but not as severe as Chinese.

12. Although **CMC** has so far been mainly text-based, the Internet is also being used for voice communication, which is technically called **Voice over Internet Protocol** (*VOIP*). The main advantage of this technology from the users' point of view is the considerably reduced price for international calls as compared with standard circuit-switched calls,

particularly in the case of PC-to-PC communication. However, the low cost comes with the trade-off of inconsistent **voice quality** due to the use of packet switching, which was designed primarily to deliver non real-time data.

Chapter 5

Globalization and Localization: Culturalization of Content and Package

✍**1.** ***Chapter 5*** examines how the globalization process is fundamentally affecting Translation, in particular with the need for **localization**. Building on Gile's (1995) concept of the message consisting of '**Content**' and '**Package,**' the authors highlight a new dimension of Translation: **culturalization** of the message. The chapter discusses the importance of language management in globalization.

✍**2. Localization** is now being applied to both the Content and Package of wide-ranging products and services to render the Message as a whole into an appropriate form in the cultural context of the Receiver. We call this process the '**culturalization**' of the Message.

✍**3.** The authors define **globalization** in relation to Translation-mediated Communication (TMC) as: 'a process to enable the Message to be adaptable to the condition that may be imposed by Receivers who do not share the same linguistic and cultural backgrounds as the Sender.'

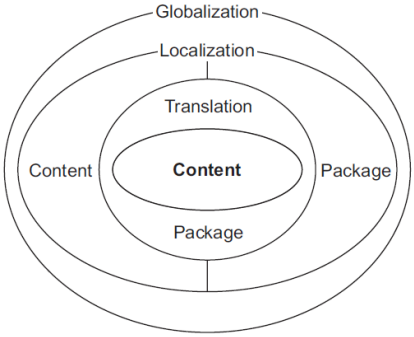
✍**4.** The term '**localization**' can be defined as 'a process to facilitate globalization by addressing linguistic and cultural barriers specific to the Receiver who does not share the same linguistic and cultural backgrounds as the Sender.'

✍**5. Web localization** has come to involve not only the **Content** of the Message but also that of the **Package—**such as the general design of the home page, the layout, the font, the color scheme, the icon design and the positions of buttons.

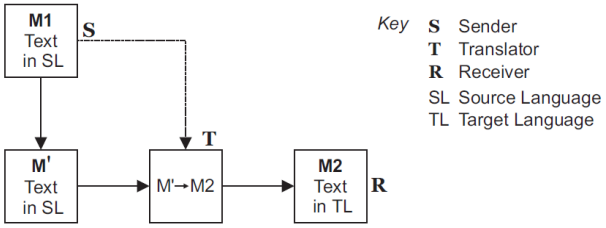
✍**6.** In **Web localization**, the term '**content management**' is used to include: ❶ localization of the Web site and ❷ maintaining the given Web site. It is therefore different from our own use of the term **Content**. In this chapter, we use **Content** with a capital **C** to mean specifically 'the words and linguistic structures' of the Message whereas '**Package**' includes any other non-textual elements and the container (medium) in which the **Content** is delivered.

✍**7.** In the context of **globalization**, TM Chas generally come to mean **Receiver-oriented** messaging in the form of **localization** and implies

that both Content and Packaging normally undergo transformation. *Figure 5.1* illustrates the relationship between **globalization**, ***localization*** and **translation**. In one sense, Translation is a core to both localization and globalization, but in another sense, without the engineering inputs of localization, globalization on the Internet is not feasible. The diagram also shows how Translation in general can be seen as more concerned with Content than Packaging whereas in localization **Packaging** is as important as the **Content**.



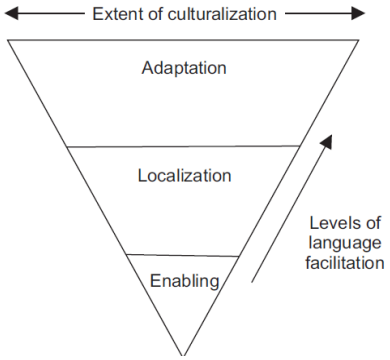
8. *Figure 5.2* illustrates TMC, focusing on the change in the **Message** [M1] transformed to [M'] by the **internationalization** process before it is translated by the Translator into [M_2]. The Message as [M'] is still in the source language, but is considerably more amenable to the ensuing translation/localization process (see also *Figure 2.3*). The **internationalization** process to convert [M_1] into [M'] now creates a new type of pre-translation work. In fact, this may remind some readers of the '**pre-editing**' process routinely applied to texts to be processed by Machine Translation (MT) to make them more 'machine-friendly' by eliminating known difficulties such as ambiguities and complexities. However, **pre-editing** for MT is carried out primarily to simplify the machine translation process, while **internationalization** (as far as its non-technical aspects are concerned) aims at human consumption and involves changing not only the **Content** but also the **Package**.



9. Microsoft applies three **incremental levels** of facilitation consisting of what they call Enabling, Localization and Adaptation as described below:

- ◈ First level: **Enabled** – users can compose documents in their own language, but the software user-interface and documentation remain in English.
- ✪ Second level: **Localized** – the user-interface and documentation are translated, but language-specific tools and content remain in English.
- ※ Third level: **Adapted** – the linguistic tools, content, and functions of the software are revised or re-created for the target market. (Brooks, 2000: 49)

✍**10.** *Figure 5.3* **Language management** with levels of *language facilitation* (Microsoft model):



✍**11.** The **first level** of facilitation as enabling does not involve Translation in its traditional sense, as the given software product is mainly adjusted at a technical level to allow inputs in the script of a given language. At **the second** level of language facilitation, the Message is converted into the Receiver's linguistic environment and a degree of cultural adjustment may be made in terms of basic features. The **third level** means that the Message is fully adapted to the Receiver's linguistic and cultural environments.

✍**12.** While the **Internet** has steadily become the mainstream communication medium in most developed countries since the mid-1990s, Japan is considered to have been slow in its adoption of the **Internet**. However, the introduction of a wireless Internet service '**i-mode**' seems to have finally, and rather unexpectedly, launched the country into the **Internet** era. According to its developer, **i-mode** terminals are deliberately designed to retain the appearance of a phone, with the built-in **Internet** access almost hidden as part of the telephone functionality (Matsunaga, 2000). To access the Internet, the user needs only to press the '**i-mode**' button.

PART III

Emerging Domains of Translation Practice

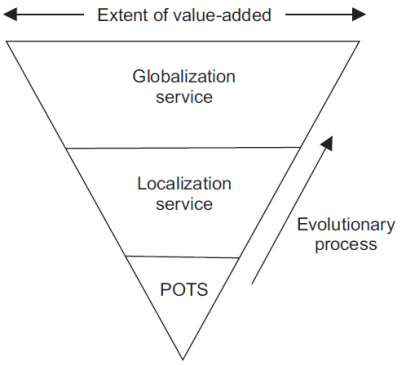
💻***Part III*** moves to the coalface of Translation practice by focusing particularly on teletranslation and teleinterpretation.

Chapter 6
Teletranslation

✍**1.** ***Chapter 6*** observes how **teletranslation** is operating and advancing. It highlights key emerging trends towards mature teletranslation.

✍**2.** The authors define '**teletranslation**' to mean:
⌘ translation operated via the Internet and
☆ translation of Internet-related content.

✍**3.** *Figure 6.1* Emergence of the **teletranslation** industry:



✍**4.** *Figure 6.1* represents the emerging picture of the **teletranslation** industry in which the electronic network is used to facilitate the customer interface, distribution of text and sometimes the translation function itself, as in the case of online translation services. Furthermore, the main translation work undertaken is also characterized by its direct link to the digital media. **Teletranslation** can be seen as evolving in response to the need for a language-processing capability to deal with the new types of Messages being developed that are specific to the Internet environment and, at the same time, leveraging the advantage afforded by the worldwide information infrastructure based on the Internet.

✍**5.** In the simplest way, **CMC** can be depicted as a mode of communication with a computer placed between the **Sender** and the **Receiver** carrying the **Message**.

✍**6.** *Figure 6.3* illustrates how the **teletranslation** processing function is extended to M1 by way of internationalization or in some other consultative manner. This process may sometimes involve the Receiver,

as shown by a dotted line in the figure. Also storage and processing functions are linked in cases such as TM tools, which may sometimes be accessed by the Sender as well. With increasingly seamless functionality of communication systems in terms of storage, transmission and processing, **teletranslation** systems will likely see these functions gradually converge. Processing and storage functionality are already converging in TM. Similarly, the transmission mechanism may also become transparent with '**always-on**' access to the Internet, as is already the case with wireless communication and is likely to be extended to other communication systems in the future.

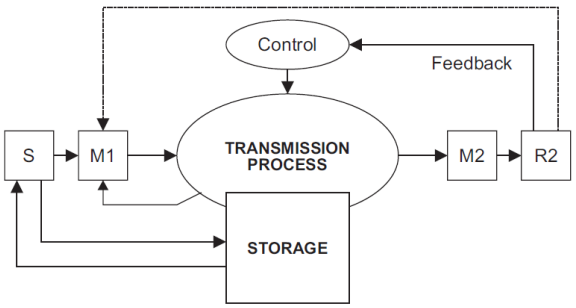


Figure. 6.3 Teletranslation Communication System

Chapter 7
Teleinterpretation

1. ***Chapter 7*** turns to remote modes of interpretation such as telephone interpreting in relation to the future development of **teleinterpretation**. A number of critical issues are discussed in the path towards **teleinterpretation**.

2. In contrast with translation, which facilitates asynchronous text-based communication, interpretation in its traditional mode deals with synchronous speech-based interactions. This fundamental difference in the mode of communication that interpretation and translation facilitate has meant a delay of the former to develop into **teleinterpretation**.

3. *Table 7.1* **Teleconference** modes in view of **remote interpreting**:

Teleconference Modes	
Type	*Description*
Audio Conference	voice only (conference call)
Video Conference	voice and facial images (video telephony)
	voice and moving images (studio- or room-based)
Audiographic Conference	voice with text and other visuals
Computer Conference	text chat sometimes combined with voice chat, including dynamic links to Web
	moving images in addition to above

✍**4.** The fundamental difference between **telephone interpreting** and other types of remote interpreting such as *videoconference* interpreting lies in the fact that telephone interpreting uses **telecommunication** as a medium to make an interpreting service available without the interpreter being present in person.

✍**5.** The situation in which **remote interpreting** may occur can be categorized as follows:

❄ The Sender and the Receiver are in the same location (face to face) with the interpreter linked via telecommunications.

❋ The interpreter is in the same location with either the Sender or the Receiver but the Sender and the Receiver are in separate locations and linked via telecommunications.

⌖ The Sender, the Receiver and the interpreter are each in separate locations and linked via telecommunications.

✍**6. Telephone** interpreting may occur in any of the above situations, whereas **videoconference** interpreting most typically occurs in situation 2. In this sense, **videoconference** interpreting is not carried out in truly telecommunication-based form, whereas telephone interpreting in situations 1 and 3 shows the characteristics similar to **teleinterpretation** where the interpreter is not physically present with either the Sender or the Receiver.

✍**7.** One major difference between the conventional form of **interpretation** and **teleinterpretation** is that in the latter the Sender and the Receiver are in different physical locations. And yet, unlike the conventional form of translation, which has always had the Sender and the Receiver in separate locations, in **teleinterpretation** all communication parties are linked via a synchronous communication mode.

✍**8.** Unlike **telephone interpreting**, the communication space, which the Sender and the Receiver share, allows interactions using text or other

visual images in addition to voice. This may mean that the interaction does **not** have to rely on spoken words alone, since other communication channels are available for visual displays of diagrams or any other images.

✍**9.** Oviatt and Cohen (1992) found that, in **telephone interpreting**, interpreters assume an independent, managerial role regarding information sequencing, including turn giving. Computer conferencing in CMC often requires a meeting **facilitator** to adopt such a role and it is easy to imagine that the role of the **teleinterpreter** would also involve facilitation. In particular, given that the Sender and the Receiver are not able to communicate directly, it will be difficult to establish and enforce protocols of turn taking during the interaction. It will be appropriate, in some cases, for the **teleinterpreter** to assume the role of mediator, thereby facilitating a smooth flow of TMC.

✍**10.** In ***face-to-face*** interpreting environments, the interpreter imparts the *intended meaning* of the Sender by combining the ***nonverbal*** and ***verbal*** elements of the Message in the source language into a verbal rendition in the target language. However, it is extremely difficult to quantify just how much information is conveyed by nonverbal cues, particularly given that in both intra and inter-lingual communication such nonverbal cues are not always consciously produced or received (Argyle, 1988).

✍**11.** In the context of *interpreting*, **videoconferencing** has often been accused of losing such *cues*, thus straining the interpreter's concentration (Kremer, 1997). On the other hand, there is a report (Mintz, 1998) that **telephone interpreting** is better in terms of **concentration** on the Message without any distraction from other channels that may be presented to the interpreter. There is a similar report on the **positive** aspect of **not** having the moving **image** of the speaker in educational settings on the Internet, as such images tend to **distract** the participants (Palloff & Pratt, 1999).

✍**12.** In summary, two problems exist for **teleinterpretation**: ❶ **bandwidth** issues, and ❷ methods of **managing** communication between teleinterpreter and the other parties to the communication. The second issue grows in complexity if we consider future situations that involve immersive virtual reality such as **HyperReality**. Another current problem is the *lack of familiarity* of conventional interpreters with the telecommunication environments that may be used for interpreting, and the need to *adapt* to situations that have not arisen in conventional interpreting.

PART IV
Future Tense

Chapter 8
Virtual Communities for Translators and Interpreters

1. ***Chapter 8*** is an examination of the Internet as a platform for *professional* developments for translators and interpreters in response to new skill and knowledge requirements. It discusses **Web-based** courses for translators with reference to case studies, and touches on future prospects for such courses for interpreters.

2. We have to keep in mind that the philosophy of management of startup (as well as established) **online companies** involves the recognition of the need to become and remain innovative in the use of technologies. In information technology, 'change is the name of the game'. As a result, new tools for telecommunication emerge almost weekly. In this climate, both the teacher and the learner of **teletranslation** and **teleinterpretation** must remain in a constant adaptive mode for two reasons:

- ongoing shifts in the delivery of multilingual support;
- the pervasive need to deal with real-time problems when engaged in synchronous communication.

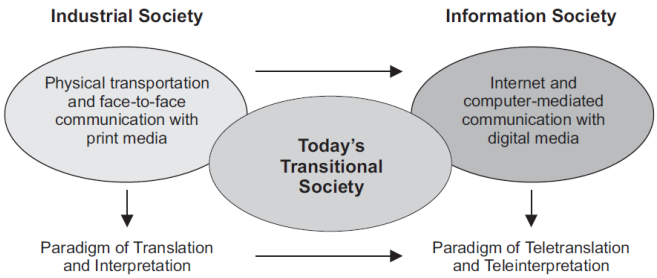
3. The second item above implies the need to **design instructional content delivery** with more than sufficient redundancy to *compensate* for transmission failures in synchronous communications, at least until the technology is fully stable.

Chapter 9
Global Information Society and the New Paradigm of Language Support

1. ***Chapter 9*** envisages the role Translation may play in the future **information society** based on extensive digital communications networks. It examines the emergence of a **new paradigm** of language support, and provides a number of future scenarios.

2. We define the **information society** as one based on an infrastructure of IT where people place greater reliance on the telecommunications system than on the physical transport system (Wang & Dordick, 1993).

3. *Figure. 9.1* A shift to information society and a new paradigm of **teletranslation** and **teleinterpretation**:



4. The role of language facilitation, as we call **teletranslation** and **teleinterpretation** in a global information society, is to serve seamlessly in the digital environment for a variety of CMC. The new dimension emerging from this role of the **teletranslator** or **teleinterpreter** may be summarized as:

- a high level of digital literacy, in particular, familiarity with given communication modes;
- an understanding of the context of the Message and the client's TMC needs;
- an understanding of wider cultural issues which concern packaging of the given Message; and
- an increasing need for subject matter specialization and commitment to ongoing professional development.

Chapter 10

New Paradigm of Translation and Interpretation

1. ***Chapter 10*** draws our argument into a vision of **teletranslation** and **teleinterpretation** as the future of translation and interpretation and highlights key issues for Translation-mediated Communication.

2. The development of online chat via interactive text has presented a potential new mode of Translation, which we have called **transterpreting**. We have suggested that **teleinterpretation** is likely to include this kind of interactive text processing, which is a hybrid between translation and interpretation. **Transterpreting** can be seen as related to the existing mode called sight translation.

3. Sight translation, which is sometimes performed by interpreters in face-to-face meetings, normally consists of reading a source-language text aloud in the target language (thus interpreting the written text in real-time) or consecutively interpreting a speech that has been read from a text.

4. The **word processor** has allowed the translation process to adopt a mode similar to **oral translation**, whereby the translator is able to input

text spontaneously even if it is not yet well formulated. In other words, **word processing** allows for relatively instantaneous production of drafts that are much more easily edited than was possible on the typewriter. This gives the translator much more room to work creatively. The main difference between this situation and **transterpreting** is that the former has the source text available at all times whereas the chat text keeps moving as it is continuously produced in real-time.

✍**5.** Also, the **transterpreter** has very little time to look up words or to go back and forth between the source and the target text. This, in fact, is a characteristic of **interpretation**, and so chat text as the Message tends to call for a kind of **synchronous** rendering more familiar to the process of **interpretation** than of **translation**. Nevertheless, both inputs and outputs are written text rather than speech, which the interpreter would normally process.

✍**6.** What happens when one **comprehends** a text is that one mentally creates a kind of world; the properties of this world may depend quite a bit on the individual interpreter's own private experiences – a reality which should account for part of the same text. As one continues with the text, the details of this world get filled in, expectations get set up which are later fulfilled or thwarted or left hanging … (Fillmore, 1977: 61)

✍**7.** According to ***Fillmore***, the message provides the reader with a '**frame**,' defined as 'any system of linguistic choices', on which the reader activates his or her own scenes of mentally created pictures.

✍**8.** As admitted by Fillmore, the term '**scene**' is used in 'maximally general sense' encompassing not only visual scenes but …any kind of coherent segment …of human beliefs, actions, experiences, or imaginings.' Using this concept, Fillmore explains the process of communication as involving 'the activation, within speakers and across speakers, of linguistic frames and cognitive scenes. Communicators operate on these scenes and frames…' (Fillmore, 1977: 66).

✍**9.** ***Vermeer*** (1992) applies Fillmore's theory to explain the translation process of nonverbal communication expressed in written texts and suggests that *distorted scenes* evoked by the translator will lead to **mistranslation**, while pointing out that the *error* can also occur in *frames*. Vermeer (1992: 288) asserts that the translator's *failure* to imagine **the scene** of a particular nonverbal behavior described in the source text will mean that the Receiver **cannot** build up his or her scene of that particular behavior.

✍**10.** Seleskovitch explains the role of the **scene** in interpretation:

> 🕮 The conceptual image that the interpreter visualizes and converts into language will similarly evoke an image in the minds of those listening to him; the image he visualizes will be colored by their own experiences, but the image may well correspond to the image they would have visualized if they had heard the original words. (Seleskovitch, 1994: 49)

✍ **11.** Given the heavy **information-processing** load that the interpreter normally has (Gile, 1995), this may mean that a second interpreter is needed to work exclusively on nonverbal conversions as a way of adjusting nonverbal cues and contextual elements. This will create a sophisticated real-time **inter-lingual** and **inter-cultural** communication facilitation whereby not only the *frames* (**Content**) but also the *scenes* (**Package**) are changed, by manipulating nonverbal cues and contextual information.

✍ **12. Key Issues for Translation-mediated Communication**

- ◈ **Issue of quality:** Quality of translation has been considered as not readily quantifiable in the sense that there is always more than one way to translate the same sentence. By comparison, because of its interactive nature, interpretation has had more immediate means of receiving user feedback. The emergence of the localization industry has had a significant influence through its efforts to quantify and benchmark the quality of translation. As a result, many translation operators are ISO (International Standardization Organization) accredited, or striving to gain such accreditation.
- ▣ **Issue of machine-assisted production:** Conventional language facilitation was entirely dependent on human efforts. This has changed drastically owing to the development of technology, which has significantly influenced the production of translation although it has not yet had the same impact on interpretation. Today's translation competence includes the proficient use of technology. For example, certain Messages created in digital environments are impossible to process without the use of technology at some point in the translation. Also, some clients are rightly or wrongly making the use of technology such as TM by the translation provider compulsory, and look on the lack of technology in the production process as lack of translation competence.

Short Answer Items & Tests

2.2 Short Answer Items

1. Sager (1993) stresses the role of Translation: 'as a commissioned task, which starts with a need for ……. and ends with a finished product.'

2. According to Kiraly (2000), translation competence refers primarily to the competence to produce ……. translations.

3. Localization is now being applied to both the Content and ……. of wide-ranging products and services to render the Message as a whole into an appropriate form in the cultural context of the Receiver. We call this process the '…….' of the Message.

4. The term '…….' can be defined as 'a process to facilitate globalization by addressing linguistic and cultural barriers specific to the Receiver who does not share the same linguistic and cultural backgrounds as the Sender.'

5. According to Kiraly (2000), translator competence refers to the ……. and ……. a translator needs in addition to translation competence.

6. Web localization has come to involve not only the Content of the Message but also that of the …….—such as the general design of the home page, the layout, the font, the color scheme, the icon design and the positions of buttons.

7. In Web localization, the term '……. management' is used to include: ❶ localization of the Web site and ❷ maintaining the given Web site.

8. In the context of globalization, TM Chas generally come to mean …….-oriented messaging in the form of localization and implies that both Content and Packaging normally undergo transformation.

9. One major difference between the conventional form of interpretation and teleinterpretation is that in the latter the Sender and the Receiver are in ……. physical locations. And yet, unlike the conventional form of translation, which has always had the Sender and the Receiver in separate locations, in ……. all communication parties are linked via a synchronous communication mode.

10. two problems exist for teleinterpretation: ❶ ……. issues, and ❷ methods of managing communication between teleinterpreter and the other parties to the communication.

2.3 Answers

1) communication	**2)** acceptable
3) Package, culturalization	**4)** localization
5) skills, knowledge	**6)** Package
7) content	**8)** Receiver
9) different, teleinterpretation	**10)** bandwidth

☙☆❧ 2.4 Tests ☙☆❧

✍ Select the best choice.

1. The main difference in modus operandi between translation and interpretation resides in the fact that ……. caters to ……. communication where all communicating parties are normally present in one physical location and communicate in real-time. By comparison, ……. facilitates ……. communication via writing with a certain time lag.

a) interpretation, asynchronous, translation, synchronous
b) translation, asynchronous, interpretation, synchronous
c) interpretation, synchronous, translation, asynchronous
d) translation, synchronous, interpretation, asynchronous

2. According to Kiraly (2000), translation ……. refers primarily to the ……. to produce acceptable translations, however one might define 'acceptable.'

a) competence, competence
b) competence, performance
c) performance, competence
d) performance, performance

3. According to Kiraly (2000), ……. competence refers to the skills and knowledge a translator needs in addition to ……. competence.

a) translation, translator
b) translator, translation
c) translation, linguistic
d) linguistic, translator

4. Dealing with digital content means that the Message is provided in a …….-readable form, which automatically facilitates the use of certain tools. MT is probably the best example, although it is hardly used as a regular 'tool' by most translation providers. Other tools include ……. Memory, terminology ……. systems, electronic dictionaries, and online databases, all of which work most efficiently with digital text.

a) human, Episodic, management
b) machine, Episodic, parsing
c) human, Translation, parsing
d) machine, Translation, management

5. ……. tools allow the translator to view a Web document without having to see the underlying HTML, XML or JavaScript code.

a) Insertion
b) Extraction
c) Transformation
d) Transference

☙✷❧ 2.5 Answer key ☙✷❧

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

Book ❸

Computers in translation:
A practical appraisal
J. Newton

3.1 Notes

Chapter 1
Introduction
John Newton

1. Systems which perform a syntactic analysis of a source text and then generate a target language rendering thereof which seeks to preserve and reconstitute its semantic and stylistic elements are described as '**machine translation**' (**MT**) systems, while those designed to facilitate human translation through providing a terminology management system, instant access to on—line dictionaries, and other utilities are referred to as '**translation tools**'.

2. Having stated that **human** translators usually score over **MT systems** in the areas of interpretation and preservation of register, it is important to stress that when it comes to spelling and terminological consistency the computer invariably outperforms the human.

Chapter 2
The story so far:
An evaluation of machine translation in the world today
Jeanette Pugh

1. In **Europe**, the impact of the **ALPAC report** was initially dramatic, but it took little more than a decade for its effects to disappear. The late 1970s witnessed a veritable explosion of MT activity in Europe, the most notable initiative being the launch by the **CEC** of the **EUROTRA** programme which has received sustained high-level funding from both the European Commission and the national authorities of all EC member states.

✍**2. Canada** is a noteworthy example of a country with an enlightened approach to machine translation. Canadian public-sector interest in MT stems from its commitment to **bilingualism**, and Canada has the claim to fame of being the first country in which an MT system (**METEO**) was put to widespread public use. The **METEO** system was originally developed at the University of Montreal and translates meteorological bulletins from English to French (Chandioux 1989).

✍**3. ALPAC** hit **American** MT very hard. The twenty-five years which followed have been referred to as the '**dark ages**' of American MT and while this assessment is perhaps an unfair reflection on the quite substantial research activities which have survived in the USA despite ALPAC, it certainly does seem an appropriate epithet when one compares the situation with the progress of events in Europe and Japan.

Chapter 3

Made to measure solutions

Annette Grimaila in collaboration with John Chandioux

✍**1.** In all real-world applications of **MT**, the **translator** is **not replaced**. In fact, he or she is the one person who must be consulted, considered and helped by the application. If the machine output is of such *low quality* or if its manipulation is so complex that the translator wastes more time revising the results than he or she would spend translating a source text, then the **usefulness** of the system is seriously **in doubt**.

✍**2.** Machine translation in its present state is far from capable of translating general texts. Human language is much **too ambiguous** for a simple machine to treat correctly and all attempts to date have been horrendously expensive if not also totally laughable. Weather bulletins, even the detailed ones prepared by *meteorologists*, are at least less ambiguous: their subject is the weather, and only the weather. Some of the remaining **ambiguities** can be circumvented by careful programming but constant adjustments are required to keep up with non-standard formulations in the source texts which are composed by human meteorologists.

Chapter 4

The Perkins experience

John Newton

✍**1.** When Peter Pym decided to explore the possibility of using MT, he already had a firm foundation on which to build: his department was

using a form of controlled English known as Perkins Approved Clear English (**PACE**). PACE was initially based on the International Language for Service and Maintenance (**ILSAM**), which in turn was based on Caterpillar Fundamental English (**CFE**).

✍ **2.** **CFE** comprised around 800 words of basic English, plus whatever technical terms were required to describe products. In 1990, the number of words in **PACE** stood at approximately 2,500, of which around 10 per cent were verbs.

✍ **3.** **PACE** is based on sound, commonsense principles: short sentences, avoidance of gratuitous synonymy (e.g. *right* is the opposite of *left*; its use in the sense of *correct* is therefore proscribed), avoidance of ellipsis, and great emphasis on clarity of expression. Founded on the principle 'one word, one meaning', the **PACE** dictionary lists and defines or exemplifies every word that is approved for use in technical publications, including articles, conjunctions, pronouns and prepositions. In the case of homographs, it specifies the parts of speech that can be used, e.g. *seal* is listed as both verb and noun, while *stroke* is listed only as a noun. The technical authors also apply a set of rules governing syntax and sentence patterns. This approach to writing grew out of a desire to convey technical information and instructions in as precise, clear and unambiguous a form as possible in the interests of safety and efficiency.

✍ **4.** Peter Pym was aware that his department's **controlled approach** to technical writing could facilitate the introduction of MT, and in March 1984 he and his colleagues established their criteria for an '**ideal**' **system**. After examining the (very few) systems that were available, they concluded that Weidner's MicroCat system matched their requirements most closely and a decision was taken to organize an operational trial using English-French. **MicroCat** is a PC-based system which processes translations in batch mode, using the transfer method.

✍ **5.** Introducing MT afforded Peter Pym a level of **control** that was previously unattainable and resulted in greater uniformity of content between source texts and their various translated versions:

> 🕮 Using MicroCat, Perkins has been able to ensure consistent terminology and to reduce translation time as well as translation costs. Using the computerized databases, Perkins can control the source and target text at all stages of publishing. Producing translation using an MT system also ensures rigorous testing and control of the source text. (Pym 1990:92)

✍ **6.** The pre-existing *systematic approach* to writing, based on continuous reappraisal, created ideal conditions for this project, as did the

enthusiastic co-operation of the technical authors and the personnel in the overseas subsidiaries. The department's relations with the latter had already been strengthened through cooperation in compiling the bilingual versions of the **PACE** dictionary. Likewise, throughout the period of the **MicroCat** test (around six months), Peter Pym had kept his colleagues in France fully briefed on developments and had sought and acted upon their advice whenever any queries had arisen concerning terminology or usage.

✍**7.** At first sight, it may appear rash to have installed MT in a department which does not have any translators on its staff, but it must be borne in mind that linguists and **post-editors** were, and still are, consulted as external resources and Peter Pym has found this arrangement to be *efficient*, *flexible* and *cost-effective*.

✍**8. Perkins'** disciplined approach to the **MicroCat** trial, which had demanded and received all the preparation normally associated with full implementation, ensured that the actual implementation of English to French went very *smoothly*, as the system was delivered with the dictionary substantially tailored to the **Perkins** environment. Another factor which maximized the *efficiency* of the **Perkins** installation overall was the decision to introduce new language pairs only when those already in use were fully operational; this enabled the subsequent implementations to benefit from the lessons learned from those which had preceded.

✍**9.** More than anything else, however, the success of the **Perkins** application was made possible by the controlled and extremely consistent nature of the source texts and by a willingness on the part of all concerned to adapt the system to the working methods and the working methods to the system.

Chapter 5

Machine translation in a high-volume translation environment

Muriel Vasconcellos and Dale A.Bostad

✍**1.** The **real test** of machine translation is whether or **not** it is *effective* in large-scale operations. These may be specialized applications, as in the case of Canada's **METEO**, or they may involve the translation of a broad variety of text types. In the latter case, the purpose of the translation will dictate the characteristics of the installation, particularly the human **post-editing** component. The **purpose** can run the gamut from publication for dissemination to '**information only**'.

✍**2.** MT has been enlisted in the service of general-purpose practical translation at **PAHO** since January 1980.

✍**3.** The use of MT has not been stabilized in **PAHO**. The new technology continues to do the lion's share of the work. The decision to use MT, which rests entirely with the terminology and translation service, is based on the following **characteristics** of the input text:

- ❄ machine readability (or optical 'scan-ability');
- ✷ complexity of format; and
- ✪ linguistic characteristics (e.g. grammar, discourse genre, need for between-the-lines interpretation, etc.).

✍**4. Post-editing** seems to be a special skill, somewhat related to traditional editing. It involves learning how to preserve as much of the machine's output as possible and '**zapping**' the text at strategic points rather than redoing it from scratch. The posteditor quickly develops a set of context-dependent techniques for dealing with the patterns produced by the machine.

✍**5.** The Air Force's Foreign Technology Division (**FTD**) has conducted three extensive surveys of machine translation over the last ten years to analyze the **effectiveness** and **use of MT** and gain insights into how to improve the product. The two most important insights coming out of the surveys are:

- ✵ speed of translation is the most important consideration for FTD analysis; and
- ⌘ the existing product, partially edited MT, is deemed satisfactory in meeting most users' translation requirements.

Chapter 6

Esperanto as an intermediate language for machine translation

Klaus Schubert

✍**1.** True or untrue, the story may well confirm that a trade as uncertain as machine translation, with its extremely long payback periods, is indeed sensitive to **prestige** considerations. It could therefore seem risky to include in a machine translation project a language like **Esperanto** which has the unmerited but undeniable quality that the mere mentioning of its name calls forth the most emotional rejections from both laymen and linguists.

✍**2. Esperanto** became associated with **computational linguistics** (which in the early decades was almost exclusively machine translation) in three stages (Schubert 1989a:26–9). These may be labelled:

❶ ‘the idea’;
❷ ‘Esperanto on equal terms’; and
❸ ‘Esperanto for its specificity’.

✍**3.** The first stage, ‘**the idea**’, had its origin in the very early years of machine translation in the late 1940s and early 1950s. After the first wishful attempts, it was soon understood that natural language is more intricate than the decoding tasks the first computers had performed well for military and intelligence applications. When natural languages turned out to be too difficult, it was suggested that something more consistent be tried, such as, for instance, Esperanto. Yehoshua Bar-Hillel put forward this suggestion in his famous state-of- the-art report of 1951 (Bar-Hillel 1951)

✍**4.** The second stage, ‘**Esperanto on equal terms**’, begins when Esperanto is actually used in computational linguistics.

⌖ First, a series of studies appear which merely investigate the feasibility of the idea,
◎ then smaller programs are written of which only a minority may have been published, and,
❁ finally, larger implementations are realized.

✍**5.** The third stage, which I term ‘**Esperanto for its specificity**’, begins with the DLT machine-translation project. Distributed language translation (DLT) is the name of a long-term research and development effort by the Dutch software company *Buro voor Systeemontwikkeling* (BSO), in Utrecht.

✍**6.** In the **DLT** system, **Esperanto** functions as the **intermediate language**. The original idea was to include in international datacommunications networks a facility that would allow each subscriber to read and to contribute messages in their own language. Potential applications are not confined to public or corporate electronic mail; they include document management, information retrieval and other functions where variable volumes of text are stored, accessed and updated in several languages. The transmission form in such a network would then be the **DLT** system’s **intermediate** language, i.e. **Esperanto**.

✍**7. Natural-language** processing, and in particular its oldest endeavour, machine translation, sometimes seems to be a never-ending struggle towards a goal that ultimately cannot be reached. In my view, the essence of the problem lies in the fact that **language is infinite**.

✍**8.** Language is an **open-ended** system. There is always more than what can be covered by explicit rules or enumerated in dictionaries or in lists of exceptions to the rules. There are no sharp borderlines, especially

between grammatical and ungrammatical utterances, between existing and non-existing words and expressions, between allowed and forbidden combinations of words or meanings, etc. All this makes language **vague**, but this vagueness is an extremely fruitful prerequisite for language to cope with the **infinite** multitude and diversity of situations that people need to communicate about. A symbol system which is not infinite in this vast and multi-layered sense is insufficient.

9. It is not sufficient to remove ambiguity at the syntactic level. More intricate and more challenging to machine translation is ambiguity at the semantic level. The **DLT** approach addresses this problem by aligning **parallel corpora** of translated texts and taking them as the basic knowledge source. DLT's '**bilingual knowledge banks**' are built from these corpora, one side of which is in Esperanto. The bilingual knowledge banks are based on the principle of extrapolation. For extrapolation, obviously, a regular language structure is a major advantage.

10. As **Esperanto**'s main and almost exclusive function is **international communication**, it has always maintained its special suitability for communication between people with radically different linguistic backgrounds and preconceptions. This condition has favoured a development on the basis of the intrinsic regularities of the language itself, rather than through imitating other languages and adopting loan patterns. Because of this pragmatic factor, Esperanto has always developed with a natural tendency towards **consistency**. This is an important asset for its function in machine translation.

11. The experience of the **DLT** machine translation project so far has shown that **Esperanto** fulfils a specific requirement in language technology: it can be used to good advantage as an intermediate language in machine translation, when fully automatic high-quality translation from the intermediate language into the target language(s) is aimed at.

Chapter 7

Limitations of computers as translation tools

Alex Gross

1. Under machine translation one finds a further distinction between batch, interactive and interlingual approaches.

- A **batch** method has rules and definitions which help it '**decide**' on the best translation for each word as it goes along. It prints or displays the entire text thus created with no help from the translator (who need not even be present but who nonetheless may often end up revising it).

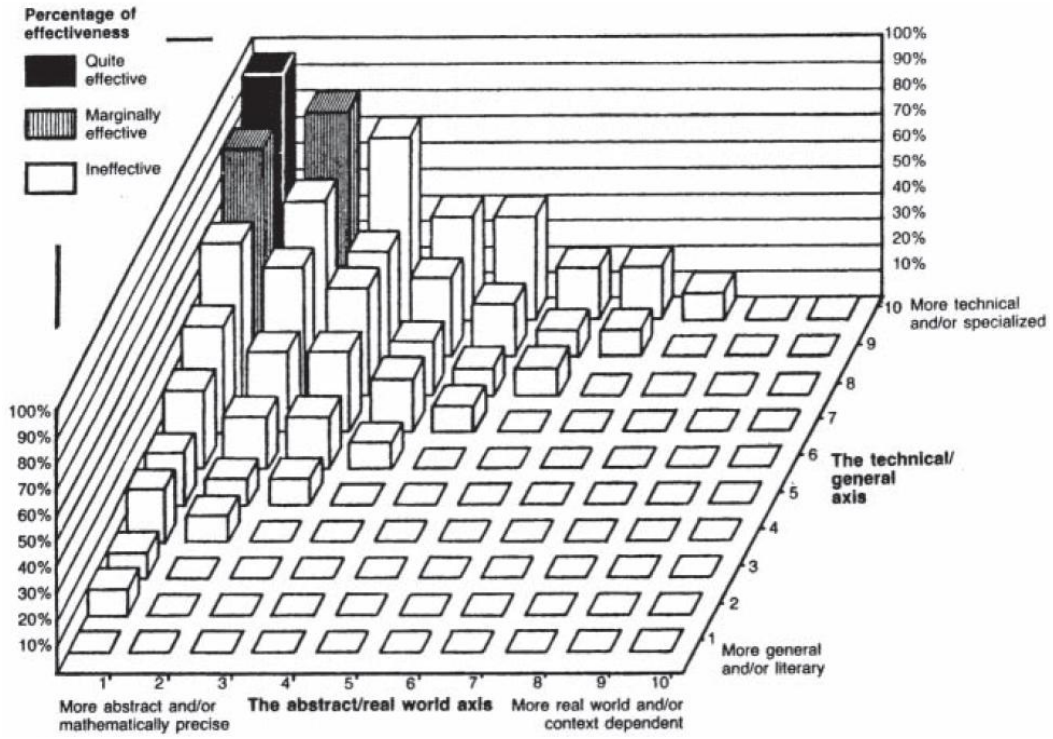
- ✯ An **interactive** system pauses to consult with the translator on various words or asks for further clarification. This distinction is blurred by the fact that some systems can operate in either batch or interactive mode.
- ❁ The so-called **interlingual** approach operates on the theory that one can devise an intermediate 'language'—in at least one case a form of Esperanto—that can encode sufficient linguistic information to serve as a universal intermediate stage—or pivot point—enabling translation back and forth between numerous pairs of languages, despite linguistic or cultural differences.

✍ **2. Batch** and **interactive** systems are sometimes also referred to as **transfer methods** to differentiate them from interlingual theories, because they concentrate on a trade or transfer of meaning based on an analysis of one language pair alone.

✍ **3.** '**Pre-editing**' means limiting the extent of vocabulary beforehand so as to help the computer. It is also used to mean simply checking the text to be translated beforehand so as to add new words and expressions to the system's dictionary.

✍ **4.** There are six important variables in any decision to use a computer for translation: speed, subject matter, desired level of *accuracy*, consistency of translation, volume and *expense*. These six determinants can in some cases be merged harmoniously together in a single task but they will at least as frequently tend to clash.

✍ **5.** The **effectiveness** of computer translation:



Chapter 8

Computerized term banks and translation

Patricia Thomas

1. Generally, little is known about **term banks** as, apart from one or two, they have not received the same press as machine translation (MT). Why is this? There seem to be three main reasons:

- First, it is only now becoming possible to buy a term bank 'off the shelf as one might a personal computer (PC) version of an MT system.
- Second, many are 'in-house' developments which are only available to specific users.
- Third, there seems to be reluctance on the part of the general public, at least in the UK, to explore the possibilities available to them from, for example, British Telecom via a telephone and a modem.

2. What sort of help can **term banks** provide? The principal functions of term banks are the storage of terms in large numbers, ease of updating, rapid retrieval and, probably most important, their standardization or

indication of preferred usage. They may provide domain classification, relationships with other terms, definitions, examples of terms in context, bibliographic references for further information and indication of copyright.

✍ **3.** What is likely to be the **structure** of **future systems**? In MT, research is being continued in two disciplines, **AI** and **CL**. It seems likely that a surface syntactic analysis of a source language (**SL**) will be underpinned by a semantic analysis, which could be used for comparison against prototypes in the form of '**frames**' or '**scripts**' in an expert system. To provide material to complete the 'slots' for frames and scripts, scanners or OCRs may be used to 'read in' texts from which an event could be inferred from partial information given; here concordancing could play a role in the provision of terms for the term bank which is an essential component in these operations.

Chapter 9

The translator workstation

Alan Melby

✍ **1.** The *functions* of a **translator workstation** can be divided into three levels (Melby 1982) as follows:

- ✷ *level one* includes word processing, telecommunications, and terminology management;
- ✪ *level two* adds text analysis, automatic dictionary look-up, and synchronized bilingual text retrieval;
- ✵ *level three* provides an interface to machine translation systems.

✍ **2.** If the purpose is simply to obtain a rough indication of the source text content, and not a careful, finished translation by human standards, then fully automatic machine translation may be in order. Raw, low-quality output which is not intended to be edited into a high-quality translation is sometimes called '**indicative**' translation.

✍ **3.** A **sublanguage text** is restricted in several ways, including vocabulary, syntax and universe of discourse. Perhaps the best known example of naturally occurring sublanguage text is weather bulletins. The **METEO** machine translation system translates Canadian weather bulletins throughout the day. In this case, human translators review the machine-translated output at a translator workstation. This workstation needs mainly *level-three functions* as only minor corrections to the raw machine-translated output are necessary.

✍ **4.** Any *true sublanguage* in which there is a large, constant flow of text is a **good** candidate for machine translation.

✍**5.** The claim of this chapter is that translator **workstations** are not just a stopgap measure to improve translator productivity until human translators are made superfluous by fully automatic high-quality machine-translation systems.

✍**6.** Both **transfer** approaches and '**interlingual**' approaches (Hutchins 1986) are based on the assumptions of modern, mainstream **generative grammar**. They treat language as a formal system in which meaning is computed compositionally from the bottom up (i.e. by combining word meanings to form larger semantic units). These word meanings are drawn from **lexicons** containing lists of predefined, distinct word senses. For each of these senses, one literal meaning is held to be more basic, while additional, metaphorical meanings are considered to be more derivative. As Lakoff (1987) has demonstrated so well, mainstream **generative grammar** is firmly based on objectivism, which is being called into question in the various branches and re-evaluations of post-modern philosophy.

✍**7.** One basic text analysis tool is a **dynamic concordance system** which indexes all the words in the document and which allows the user to request all occurrences of a word or combination of words within the document. This type of analysis may assist in the translation of a long document because it allows the translator to quickly see how troublesome terms are used in various contexts throughout the document.

Chapter 10

SYSTRAN: it obviously works but how much can it be improved?

Yorick Wilks

💣Main Notes of this article is available in Chapter 4 of the book "Wilks, Y. (2009). *Machine Translation: Its Scope and Limits*. Sheffield: Springer."

Chapter 11

Current research in machine translation

Harold L.Somers

✍**1.** The difficulties and past '**failures**' of linguistics-oriented MT point to the need for **AI** semantics-based approaches: semantic parsers, preference semantics, knowledge databases, inference routines, expert systems, and the rest of the AI techniques. (Hutchins 1986:327)

✍**2.** There is no denying the basic **AI** argument that at some stage translation involves the 'understanding' of a [source language] text in order to convey its 'meaning' in a [target language] text. (Hutchins 1986:327)

3. It is normally said that a major design advance from the first to the second generation of MT systems was the incorporation of better **linguistic theories**, and there is certainly a group of current research projects which can be said to be focusing on this aspect. This is especially true if we extend the term '**linguistic**' to include '**computational linguistic**' theories. The scientific significance of the biggest of all the MT research projects—EUROTRA—can be seen as primarily in its development of existing linguistic models, and notable innovations include the work on the representation of tense (van Eynde 1988), work on homogeneous representation of heterogeneous linguistic phenomena (especially through the idea of 'featurization of purely surface syntactic elements, and a coherent theory of '**canonical form**' (Durand et al. 1991), as well as, in some cases, the first ever wide-coverage formal (i.e. computational) descriptions of several European languages.

4. Corpus-based MT can be divided into three types, called 'memory-based', 'example-based' and 'statistics-based' translation.

5. The most '**linguistic**' of the corpus-based approaches is '**memory-based translation**' (Sato and **Nagao** 1990): here, example translations are used as the basis of new translations. The idea—first suggested by Nagao (1984)—is that translation is achieved by **imitating** the *translation of a similar example* in a database. The task becomes one of matching new input to the appropriate stored translation. In this connection, a secondary problem is the question of the most appropriate means of storing the examples.

6. Advantages of this system are ease of modification—notably by changing or adding to the examples—and the high quality of translation seen here again, as above, as a result of translations being established a priori rather than compositionally. The **major disadvantage** is the great deal of computation involved, especially in matching partial dependency trees.

7. Example-based translation: A similar approach which overcomes this major demerit has been developed quite independently by two groups of researchers at ATR in Japan (Sumita et al. 1990), and at UMIST in Manchester (Carroll 1990). In both cases, the central point of interest is the development of '**distance**' or '**similarity**' measures for sentences or parts of sentences, which permit the input sentence to be translated to be matched rapidly against a large corpus of existing translations. In Carroll's case, the measure can be '**programmed**' to take account of grammatical function words and punctuation, which has the effect of making the algorithm apparently sensitive to syntactic structure without

actually parsing the input as such. While Sumita et al.'s intention is to provide a single correct translation by this approach, Carroll's measure is used in an interactive environment as a translator's aid, selecting a set of apparently **similar sentences** from the **corpus**, to guide the translator in the choice of the appropriate translation. For this reason, spurious or inappropriate selections of examples can be tolerated as long as the correct selections are also made at the same time.

- **8. Statistics-based approaches**: Other corpus-based approaches have been more overtly statistical or mathematical. The most notable of these is the work at **IBM** (Brown et al. 1988a,b, 1990). These researchers, encouraged by the success of statistics-based approaches to speech recognition and parsing, decided to apply similar methods to translation. Taking a huge corpus of bilingual text available in **machine-readable** form (3 million sentences selected from the Canadian Hansard), the probability that any one word in a sentence in one language corresponds to zero, one or two words in the translation is calculated. The glossary of word equivalences so established consists of lists of translation possibilities for every word, each with a corresponding probability. For example, the translates as le with a probability of 0.610, as la with probability 0.178, and so on. These probabilities can be combined in various ways, and the highest-scoring combination will determine the words which will make up the target text. An algorithm to get the target words in the right order is now needed. This can be calculated using rather well-known statistical methods for measuring the probabilities of word-pairs-triples, etc.

Short Answer Items & Tests

3.2 Short Answer Items

1. As Esperanto's main and almost exclusive function is ……., it has always maintained its special suitability for communication between people with radically different linguistic backgrounds and preconceptions.

2. The bilingual knowledge banks are based on the principle of ……. .

3. The so-called ……. approach operates on the theory that one can devise an intermediate 'language'—in at least one case a form of Esperanto—that can encode sufficient linguistic information to serve as a universal intermediate stage—or pivot point—enabling translation back and forth between numerous pairs of languages, despite linguistic or cultural differences.

4. Batch and interactive systems are sometimes also referred to as ……. methods to differentiate them from interlingual theories, because they concentrate on a trade of meaning based on an analysis of one language pair alone.

5. Both transfer approaches and 'interlingual' approaches (Hutchins 1986) are based on the assumptions of modern, mainstream ……. grammar.

6. One basic text analysis tool is a dynamic ……. system which indexes all the words in the document and which allows the user to request all occurrences of a word or combination of words within the document.

7. Corpus-based MT can be divided into three types, called 'memory-based', '…….-based' and '…….-based' translation.

8. The most 'linguistic' of the corpus-based approaches is '…….-based translation' (Sato and Nagao 1990).

3.3 Answers

1) international communication	**2)** extrapolation
3) interlingual	**4)** transfer
5) generative	**6)** concordance
7) example, statistics	**8)** memory

☙☆❀ 3.4 Tests ☙☆❀

✍ Select the best choice.

1. ALPAC hit American MT very hard. The twenty-five years which followed have been referred to as the '…….' of American MT and while this assessment is perhaps an unfair reflection on the quite substantial research activities which have survived in the USA despite ALPAC, it certainly does seem an appropriate epithet when one compares the situation with the progress of events in Europe and Japan.

a) golden ages b) dark ages
c) depression period d) critical period

2. Peter Pym was aware that his department's ……. approach to ……. writing could facilitate the introduction of MT, and in March 1984 he and his colleagues established their criteria for an 'ideal' system.

a) technical, proscribed b) prescriptive, technical
c) prescribed, controlled d) controlled, technical

3. Esperanto became associated with computational linguistics (which in the early decades was almost exclusively machine translation) in three stages (Schubert 1989). These may be labelled: 'the …….'; 'Esperanto on ……. terms'; and 'Esperanto for its …….'.

a) ideal, specific, equality b) design, diverse, equality
c) idea, equal, specificity d) devise, assorted, specificity

4. In the DLT system, ……. functions as the ……. .

a) Natural-language, bilingual knowledge bank
b) Esperanto, intermediate language
c) Esperanto, bilingual knowledge bank
d) Natural-language, intermediate language

5. It is not sufficient to remove ambiguity at the ……. level. More intricate and more challenging to machine translation is ambiguity at the ……. level. The DLT approach addresses this problem by aligning parallel corpora of translated texts and taking them as the basic knowledge source. DLT's 'bilingual knowledge banks' are built from these corpora, one side of which is in ……. .

a) syntactic, semantic, Esperanto b) semantic, syntactic, Esperanto
c) pragmatic, semantic, interlingua d) pragmatic, syntactic, interlingua

☙✷❧ 3.5 Answer key ☙✷❧

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

Book ❹

Readings in Machine Translation

S. Nirenburg, H. Somers, & Y. Wilks

4.1 Notes

PART I: HISTORICAL

Introduction

Sergei Nirenburg

1. From the **1960s** on, MT was, in fact, often used to apply contemporary linguistic theories, but the systems that were directly inspired by a particular linguistic theory were usually seldom comprehensive or broad-coverage.

2. The **stochastic** approach to MT had its beginnings not in the late **1980s**, as many believe, but thirty years earlier.

3. In the **1960s**, the field gradually became much more **method-oriented**, and many (though definitely not all) projects, while paying lip service to the practical needs of **MT**, would concentrate much more on applying and testing a variety of linguistic (e.g., syntactic) and computational linguistic (e.g., parsing) theories within the framework of **MT**. The pendulum would swing once again in the late **1980s**, when the renewed emphasis on results and system evaluation in competition would bring back the engineering methods and attitudes familiar from the early days of **MT** and often quite detached from the knowledge accumulated in linguistics.

Chapter 1

Translation

Warren Weaver

1. ***Warren Weaver*** energized the early MT research, not least through his influence on the funding priorities at the National Science Foundation of the United States. Thus, among other recipients of early grants to carry out experiments in non-numerical applications of computing was *Andrew Booth* of Birkbeck College of the University of London, who concluded, in late 1947, that MT was a prime area for such an endeavor.

✍**2.** There is no need to do more than mention the obvious fact that a **multiplicity of languages** impedes **cultural interchange** between the peoples of the earth, and is a serious deterrent to international understanding.

Chapter 2
Mechanical Translation
A. D. Booth

✍**1.** The contribution by ***Booth*** in this collection describes some of his early **experimental settings** and ideas about **MT**. It is a very interesting document in that the reader should realize that the work described was truly trail-blazing and pioneering. There was no paradigm of MT research in existence yet, and even though Booth does not present his work in a paradigmatic mode, some tacit assumptions about it are interesting to note.

✍**2.** Richens pointed out that, with certain limitations, an adequate or passable **translation** of a foreign language text would result from the following operation:

- ⌘ The memory contains a stem (or root) dictionary and an ending dictionary.
- ☸ The stem dictionary consists of a relatively few entries of general semantic utility plus a vocabulary specific to the subject of the translation.

Chapter 3
The Mechanical Determination of Meaning
Erwin Reifler

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

✍**2.** Thus, the following observation about the process of translation sets up the overall view of MT as a process of **ambiguity resolution**. ''A complete message contains information that, together with a certain number of unsymbolized situational criteria, enables the human hearer, reader, or translator to select the intended meanings from the multiple potential meanings characterizing its constituents.''

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

✍**4.** The determination of **intended meaning** depends not only on the semantic peculiarities of the source language, but on the semantic peculiarities of the target language as well! As already mentioned, our problem is multiple meaning in the light of source-target semantics. If, for instance, we want to translate the English sentence, ‘‘*He is an ass*,’’ into Chinese, we must discover whether the Chinese word for ‘‘*ass*’’ can be used as a contemptuous expression denoting a stupid human being.

✍**5.** As a matter of fact, it cannot be so used, and therefore a **literal translation** would be completely **unintelligible**. Another Chinese word meaning something like ‘‘stupid’’ or ‘‘foolish’’ has to be substituted or else the English sentence has to be expressed in a completely different way according to the idiomatics of the Chinese language. Of course, most of the present-day MT systems do not attempt to resolve this type of problem dynamically, and typically are only capable of doing this (or even considering this as a problem!) if the appropriate reading is listed among the senses in the transfer dictionary.

✍**6.** Another interesting find is the following early statement concerning, essentially, the issue of selectional restrictions: From among multiple **non-grammatical meanings** the translation mechanism will extract the intended meaning by determining the **non-grammatical meaning** in which two or more syntactically correlated source forms coincide. For example, in *Er bestand die Prufung* (he passed the examination) the memory equivalent of *bestand* will be accompanied by a number of distinctive code signals, each indicative of one of its multiple non-grammatical meanings. One of these code signals will be identical with a code signal accompanying the memory equivalents of all substantives which, as objects of *bestand*, ‘‘pinpoint’’ the intended meaning of the latter as one best translated by English ‘‘passed.’’

Chapter 4

Stochastic Methods of Mechanical Translation

Gil King

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

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

Chapter 5

A Framework for Syntactic Translation

Victor H. Yngve

2. The contribution by ***Victor Yngve*** belongs to the wave of MT efforts that followed the initial experimentation. It represents more mature research activities that led the field to deeper and more comprehensive descriptions of the requirements and approaches to MT. Yngve’s paper enumerates types of clues for source text analysis, anticipating the central issues of the area of natural language parsing. It also introduces an influential discussion of the ‘‘100%’’ vs. ‘‘95%’’ approaches to MT:

3. The six types of [analysis] clues are:

❶ The field of discourse.

❷ Recognition of coherent word groups, such as idioms and compound nouns.

❸ The syntactic function of each word.

❹ The selectional relations between words in open classes, that is, nouns, verbs, adjectives, and adverbs.

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

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

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

Chapter 6

The Present Status of Automatic Translation of Languages

Yehoshua Bar-Hillel

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

✍ **2. On the 95 percent approach:**

It is probably proper to warn against a certain tendency which has been quite conspicuous in the approach of many MT groups.

These groups, realizing that **FAHQT** *[Fully automated, high-quality MT]* is not really attainable in the near future so that a less ambitious aim is definitely indicated, had a tendency to compromise in the wrong direction for reasons which, though understandable, must nevertheless be combated and rejected. Their reasoning was something like the following: since we cannot have **100%** automatic high-quality translation, let us be satisfied with a machine output which is complete and unique, i.e., a smooth text of the kind you will get from a human translator (though perhaps not quite as polished and idiomatic), but which has a less than **100%** chance of being correct. I shall use the expression ‘‘**95%**’’ for this purpose since it has become a kind of slogan in the trade, with the understanding that it should by no means be taken literally. Such an approach would be implemented by one of the two following procedures: the one procedure would require to print the most frequent target-language counterpart of a given source-language word whose ambiguity has not been resolved by the application of the syntactical and semantical routines, necessitating, among other things, large scale statistical studies of the frequency of usage of the various target renderings of many, if not most, source-language words; the other would be ready to work with syntactical and semantical rules of analysis with a degree of validity of no more than 95%, so long as this degree is sufficient to insure uniqueness and smoothness of the translation.

3. On statistics and MT:

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

4. On context and ambiguity resolution:

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

gotten beyond the barest and vaguest outlines—is still powerless to determine the meaning of pen in our sample sentence within the given paragraph? *[Here Bar Hillel refers to his famous example of the text ''Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy.'' Where the word ''pen'' cannot be disambiguated between the writing implement and enclosure senses without the use of extralinguistic knowledge about the typical relative sizes of boxes and pens (in both senses).]*

Chapter 7

A New Approach to the Mechanical Syntactic Analysis of Russian

Ida Rhodes

1. The contribution by ***Ida Rhodes*** is a very well reasoned and meticulously argued presentation of results of practical MT system development, with a realistic perspective on the complexities of the task at hand. First of all, Rhodes forcefully describes the objective obstacles in the path of a translator, even a human translator, let alone a computer program. She elegantly concludes that

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

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

Chapter 8

A Preliminary Approach to Japanese-English Automatic Translation

Susumu Kuno

✍**1.** The contribution by ***Susumu Kuno*** describes a method for Japanese–English MT, with an original Japanese **segmentor** and **syntactic analysis** following the method of Rhodes. At the time of publication, the method was **not** yet implemented in a computer system, but it describes the first attempt at solving a very important problem in processing Asian languages (and other languages with no breaks between words) that has achieved some prominence in the late 1980s and in the 1990s.

✍**2.** The results of preliminary manual testing of **automatic segmentation** on the basis of a ‘‘find the longest matching dictionary item’’ operation followed by ‘‘predictive testing’’ has given reason to believe that this program will provide a practical basis for the analysis of running kana-kanji text. Thirty-nine distribution types for Japanese have thus far been recognized, but no exhaustive classification of dictionary and auxiliary items into these types has been attempted. In particular need of further study are the problems of homographs and missing words.

Chapter 9

On the Mechanization of Syntactic Analysis

Sydney M. Lamb

✍**1.** The contribution by ***Sydney Lamb*** seems to be a prolegomenon to the currently very fashionable studies devoted to inducing **syntactic grammars** from corpora and will give a historical perspective for this type of activity.

✍**2.** There are three (and only three) types of **hierarchal relationships** existing among the **structural units** of language. They are:

❶ that of a class to its members (e.g., vowel: /a/, noun: **B**oy);
❷ that of a combination to its components (e.g. /boy/:/b/, <men and women>: <women>); and
❸ that of an *eme* and its *allos* (e.g., /t/:[t′]).

✍**3.** These relationships may be called hierarchical because in each of them there is one unit which is in some way on a higher level than the others. There is a ***fourth type*** of hierarchical relationship, but it is not present within the structure of a language. It is that of a type to its tokens, and it exists as a relationship of the language to utterances or texts. Any unit of a linguistic structure is a type with relation to tokens, i.e., occurrences, of it in texts.

Chapter 10

Research Procedures in Machine Translation

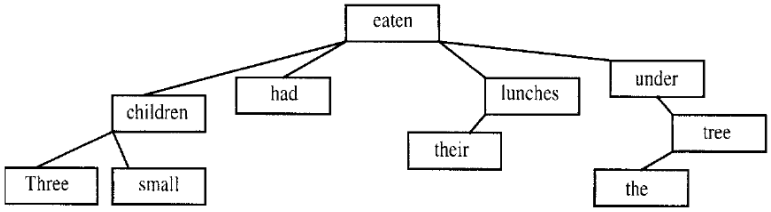
David G. Hays

✍**1.** The contribution by ***David Hays*** is mostly interesting for its acute methodological observations concerning the research tasks to be carried out by MT developers. He states:

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

✍**2. Morphology** has to do with the analysis of words and **forms** of words. In some but not all languages the word forms that occur in text can be subdivided into repetitive fragments; that is, relatively few fragments combine and recombine in many ways to yield a large vocabulary of forms. In an MT system it is economical to avoid storing repetitive data if they can be reconstructed by a simple program from a smaller base; hence storage of **fragments** instead of full forms is usually advocated by system designers.

✍**3.** According to **dependency theory**, a partial ordering can be established over the occurrences in a sentence. One occurrence is independent; all the others depend on it, directly or indirectly. Except for the independent occurrence, every occurrence has exactly one **governor**, on which it depends directly. The diagram of relations among occurrences in a sentence is a tree, an example of which is given in *figure 10.1*:



Chapter 11

ALPAC: The (In)Famous Report

John Hutchins

✍**1.** The **ALPAC** report has exerted monumental influence on the development of MT in the U.S.. It is very important for the present-day MT researcher to understand what **ALPAC** actually said because what

usually trickles down the collective memory is only the **extra-scientific** consequences of its publication, most of all the steep drop in the levels of funding of MT in the US after **ALPAC**'s publication.

✍**2.** **ALPAC** began by asking whether, with the overwhelming predominance of English as the language of scientific literature (76% of all articles in 1965), it ''might be **simpler** and more **economical** for heavy users of Russian translations to learn to read the documents in the original language.'' Studies indicated that this could be achieved in 200 hours or less, and ''an increasing fraction of American scientists and engineers have such a knowledge,'' and it noted that many of the available opportunities for instruction were underutilized.

✍**3.** On **quality**, ALPAC stressed that it must be appropriate for the needs of requesters: ''flawless and polished translation for a user-limited readership is wasteful of both time and money.'' But there were no reliable means of measuring quality, and for this reason ALPAC set up an evaluation experiment.

✍**4.** On **speed**, ALPAC saw much room for improvement: scientists were complaining of delays; the most rapid service (from JPRS) was 15 days for 50 pages; the NSF translation of journals ranged from 15 to 26 weeks; documents sent to outside contractors by the US Foreign Technology Division were taking a minimum of 65 days; and when processed by the FTD's MT system, they were taking 109 days.

✍**5.** On **cost**, ALPAC considered what government agencies were paying to human translators and this varied from $9 to $66 per 1000 words.

Chapter 12

Correlational Analysis and Mechanical Translation

Silvio Ceccato

✍**1.** The contribution by ***Silvio Ceccato*** is one of the most original ones in this volume. The famous Italian linguist presents a study elegant in style and intriguing in substance; among other reasons, this is because the author does not seem to be influenced, to any significant degree, by the MT scholarship that had been accumulated by the time this contribution appeared. While this might be considered a drawback, it also leads to an original point of view that will help us to present the MT scene as a complex and diverse phenomenon that it was. Here are some of Ceccato's opinions.

✍**2.** Echoing Rhodes' position concerning **MT glossaries**, Ceccato avers that ''the entrepreneurs of mechanical translation must have been unpleasantly surprised for grammar, as it was conceived for men, is not

immediately applicable to machines.'' He explains it in an idiosyncratic way, saying that computational grammars are not conceived as links between morphology and semantics.

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

Chapter 13

Automatic Translation:

Some Theoretical Aspects and the Design of a Translation System

O. S. Kulagina and I. A. Mel'cuk

✍**1.** The contribution by ***Kulagina and Mel'cuk*** is a bold and surprisingly modern programmatic statement about how one should understand the problem of MT and its ''**ecology**.'' In their own words:

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

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

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

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

> Understanding the ''linguistic'' meaning of a text does not guarantee the ability to process this text correctly: ''**linguistic**'' meaning and ''situational''

content (the state of affairs) are quite different things not always linked by a unique (one-to-one) correspondence. The right translation is possible only if the extralinguistic situation is rightly understood.

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

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

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

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

Chapter 14

Mechanical Pidgin Translation

Margaret Masterman

✍**1.** A similar topic is central to the article selected from the writings of ***Margaret Masterman***, an MT researcher and teacher of many other luminaries in MT and AI, including Martin Kay and Yorick Wilks:

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

✍**2.** Masterman and her colleagues researched the **semantic thesaurus** in some detail, and it might be said that that was the original work concerning **semantic interlinguas** (as opposed to syntactic ones like the one suggested by *Vauquois*). This work found further development, for instance, in the work of Sparck Jones and Wilks.

✍**3.** The paper selected for this collection describes a method of automatically transforming results of low-quality word-for-word MT (with a morphological analyzer!) into a readable form, essentially by carrying out feature transfer between source and target languages. The paper calls for more attention to what the author calls ‘‘**bits of information**’’ and we would call grammatical morphemes and closed-class lexical elements of a language. The good example of how much these elements contribute to the understanding of the meaning of text is, as Masterman mentions, a text like Lewis Carroll’s ‘‘Jabberwocky,’’ in which all open-class lexical items are not English, while all the closed class items are.

Chapter 15

English-Japanese Machine Translation

S. Takahashi, H. Wada, R. Tadenuma, and S. Watanabe

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

PART II

THEORETICAL AND METHODOLOGICAL ISSUES

Chapter 16

Automatic Translation and the Concept of Sublanguage

J. Lehrberger

1. It is not known how many **sublanguages** exist in a given language. They are not determined a priori but emerge gradually through the use of a language in various fields by specialists in those fields. They come to our attention when people begin to refer to ‘‘the language of sports-casting,’’ ‘‘the language of biophysics,’’ etc. A ***grammatical sentence*** in a Sublanguage of English may **not** be grammatical in standard English even though the text in which the sentence occurs is still said to be ‘‘in English.’’ When we speak of ‘‘the language as a whole’’ we include all such texts, thus it seems that a grammar of the language as a whole must describe all the **sublanguages** in it—certainly no mean task.

2. Many of the sentences of a **Sublanguage** of L are considered ‘‘**standard L**’’; the percentage varies within each **Sublanguage**. And those sentences that are not so considered can be paraphrased in standard L (Check reservoir full⇔Check to ensure that the reservoir is full). This suggests that the standard language may be useful in describing the way a Sublanguage fits into the language as a whole. Furthermore, sublanguages overlap and their interrelations form a part of the description of the language as a whole. A language is not simply a union of sublanguages, but a composite including many sublanguages related to varying extents lexically, syntactically and semantically.

Chapter 17

The Proper Place of Men and Machines in Language Translation

Martin Kay

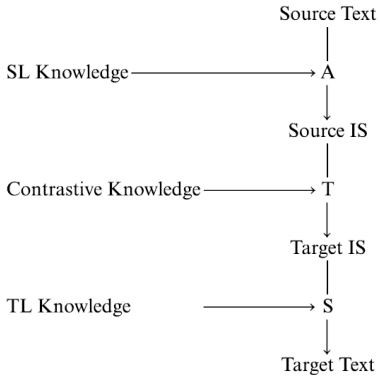
1. The translator’s amanuensis will not run before it can walk. It will be called on only for that for which its masters have learned to trust it. It will not require constant infusions of new ad hoc devices that only expensive vendors can supply. It is a framework that will gracefully accommodate the future contributions that **linguistics** and **computer science** are able to make. One day it will be built because its very modesty assures its success. It is to be hoped that it will be built with taste by people who understand languages and computers well enough to know how little it is that they know.

Chapter 18

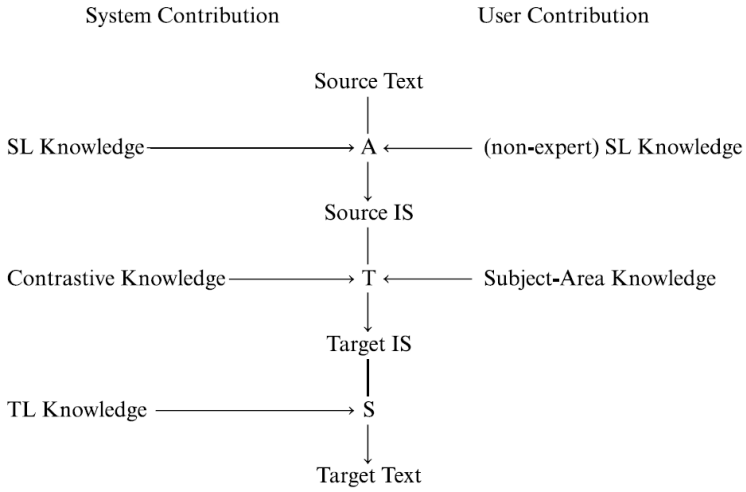
Machine Translation as an Expert Task

Roderick L. Johnson and Peter Whitelock

✍**1. A Model of Translation:** The basic model we propose, in over-simplified form, is the familiar transfer scheme shown in *figure 18.1*. The idea is that some analysis device A applies SL knowledge to a source text to produce a source internal structure IS; a transfer device T applies contrastive knowledge to the source IS to produce a target IS; and finally a synthesis device S applies **TL knowledge** to the target IS to produce a target text. In addition (not shown in the figure) all three of ***SL knowledge***, contrastive knowledge, and **TL knowledge** may be enhanced by text-type knowledge. In practice, as we all know, this model, even when enriched by text-type knowledge, is pathetically inadequate. For it even to have a chance of being useful, we should have to require that all of S, T, and A be total and functional. In practice, we know that this is unlikely ever to be the case with natural text.



✍**2.** The model we propose is **intermediate** between the pre-editing and interactive styles of MT. If the **machine** is to behave functionally as far as possible like a **human translator**, then we would like to free the user from any need to know about the target language, so that the machine has to be a TL and a contrastive expert, as well as having text-type knowledge built in. On the other hand, while we anticipate that the system will be more or less deficient in knowledge of the user's SL and in subject-area knowledge, we assume that these deficiencies can be remedied in consultation with a (SL) monolingual operator.



Chapter 19

Montague Grammar and Machine Translation

Jan Landsbergen

1. There are three problems with using **Intentional Logic** as an **interlingua**:

- The first problem is that a **meaning representation** in Intentional Logic may require a more detailed meaning analysis than is needed for translation purposes, because for translation we are mainly interested in equality of meanings. This problem is **solved** by using semantic derivation trees as interlingual meaning representations, in which the unique names of basic meanings and meaning rules serve exactly to express the equality of meaning of basic expressions and syntactic rules, respectively.
- The second problem is that expressions of **Intentional Logic** only convey the **meaning** in the strict model-theoretical sense. Semantic derivation trees indicate in addition the way in which the meaning is derived. They may also be used to convey other information than the meaning. If two basic expressions or two syntactic rules (of the same language) have the same meaning, but differ in some other aspect which is relevant to translation, we may assign different names to the corresponding basic meanings and meaning rules.
- The **solution** of the third problem, the **subset problem**, has been the main motivation for the isomorphic grammar approach. If the grammars of the source and the target language are isomorphic, each interlingual expression generated by the analysis component can be processed by the generation component.

Chapter 20

Dialogue Translation *vs.* Text Translation—Interpretation Based Approach

Jun-ichi Tsujii and Makoto Nagao

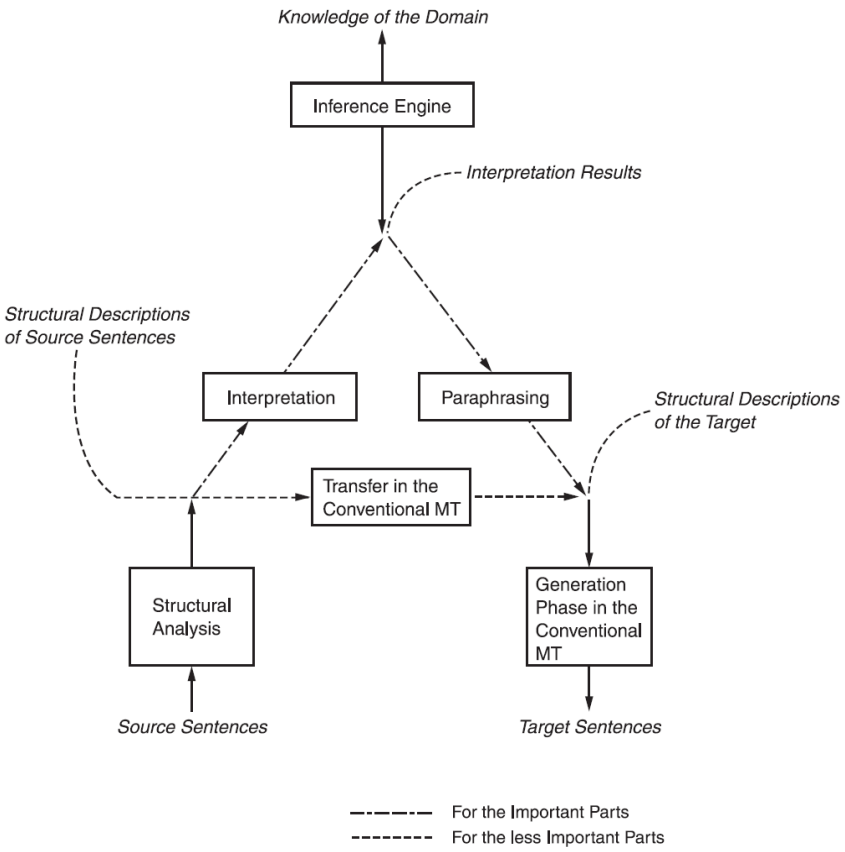
✍ **1.** We can summarize the differences of environments in which these two types of systems might be used as follows.

⌘ **Clear Definition of Information**: In certain types of dialogue translations, we can define rather clearly what information should be transmitted from source sentences to target translations, while we generally cannot in textual translation. By certain types of dialogues, we mean here the dialogues such as dialogues for hotel reservation and conference registration which are currently picked up by the ATR research group, dialogues between patients and doctors tried by the CMU group ([Tomita86]), etc.

☸ **Active Participation of Speakers and Hearers**: In most application environments of textual translation systems, they are supposed to be used by professional translators. We cannot have the writers of texts at the time of translation, the persons who prepare texts and really want to communicate something through the texts. The actual readers of translated texts are not available, either, at the time of the translation, who really want to get messages or information encoded in the texts.

✍ **2.** In **dialogue** translation, we have both the speakers (the senders of messages) and the hearers (the receivers of messages) at the time of translating messages. These two differences make, we claim, dialogue translation systems more feasible in actual translation environments, if they are properly designed for taking these advantages.

✍ **3.** *Figure 20.2* shows a schematic view of a system which translates **dialogues** in a certain restricted domain. The translation system knows in advance what kinds of **information** or **concepts** are important for the natural flow of dialogues in that specific task domain, and also knows a set of surface linguistic expressions which may convey such **important information**. By using these kinds of knowledge, the system should be able to distinguish the parts which convey important informational contents, extract them and relate them to the representations of the explicit understanding layer.

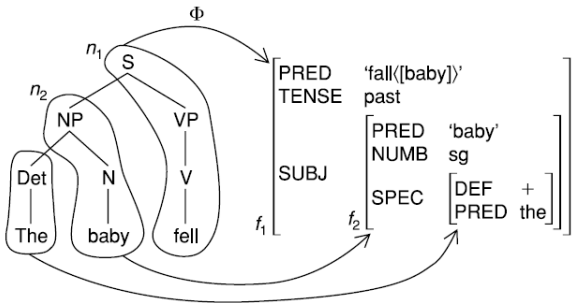


Chapter 21

Translation by Structural Correspondences

Ronald M. Kaplan, Klaus Netter, Ju¨rgen Wedekind, and Annie Zaenen

1. The formal picture developed by Kaplan and Bresnan, as clarified in Kaplan (1987), is illustrated in the following structures for sentence (1) (*figure 21.1*). The c-structure appears on the left, the f-structure on the right. The c-structure-to-f-structure correspondence, Φ, is shown by the linking lines. The correspondence Φ is a many-to-one function taking the S, VP, and V nodes all into the same outermost unit of the f-structure, f_1.



Chapter 22

Pros and Cons of the Pivot and Transfer Approaches in Multilingual Machine Translation

Christian Boitet

✍**1.** The **pivot approach** seems best suited to the construction of multilingual M(A)T systems, for obvious reasons of minimality and economy. The idea is to translate the input text into a pivot language, and then from this pivot into the target language. In a multilingual setting with ***n*** languages, only ***n*** analyzers and ***n*** generators have to be constructed, comprising ***2n*** grammars and ***2n*** dictionaries (which give monolingual information and translations into or from the pivot lexicon).

✍**2.** In the **transfer approach**, there is the same number of analyzers and generators, but ***n(n-1)*** transfers must be added. They transform source interface structures into target interface structures, using ***n(n-1)*** transfer grammars and transfer dictionaries. If the interface structures contain a deep enough level of linguistic description, the transfer grammars are very small: the transfer dictionaries represent the bulk of the cost of the ***n(n-1)*** transfers, they may be large, and they are more difficult to construct than monolingual dictionaries.

Chapter 23

Treatment of Meaning in MT Systems

Sergei Nirenburg and Kenneth Goodman

✍**1.** Most recently Brown et al. (1988) report on experiments with a **statistical approach** to machine translation which ‘‘. . . eschews the use of an intermediate mechanism (language) that would encode the ‘meaning’ of the source text.’’ The contention in this approach is that ‘‘. . . translation ought to be based on a complex glossary of correspondence of fitted locutions’’ and more fully, *translation* can be somewhat naively regarded as a **three stage** process:

- ☸ Partition the source text into a set of fixed locutions.
- ☸ Use the glossary plus contextual information to select the corresponding set of fixed locutions in the target language.
- ✠ Arrange the words of the target fixed locutions into a sequence that forms the target sentence.

✍ **2.** In other words, language in this approach is treated not as a productive system but as a fixed and unproductive set of canned locutions. The applicability of an MT system built according to this approach is restricted to the cases where there are vast textual corpora of translation equivalents.

✍ **3.** But even when such materials are available, completely uninterrupted comparison will lead to **errors** simply because the human translators who produced the translations in the corpus in the first place do not translate **word-for-word** or even **sentence-for-sentence**. The meaning expressed by a lexical unit in the source language can be rendered as an affix or as a syntactic construction in the target language. ***Nagao*** (1989:6–7) writes:

> . . . although they are infrequently used in European languages, in Japanese there are many words of respect and politeness which reflect the social positions of the speakers, as well as distinctly male or female expressions which lie at the heart of Japanese culture. These are factors which must be considered when translating between Japanese and European languages. . . . Even if those factors are not explicitly expressed in the target language, they should be inferable from the context, from the psychological state of the speaker, or from the cultural background of the language. It will be difficult for a purely statistical system to detect such phenomena.

✍ **4.** Translation is a relation between texts in the source and target languages, such that the invariant between them is meaning. In other words, translation is rendering a set of meanings realized in a source language using the realization means of a target language.

✍ **5. MT** deals with **expository texts**, where the artistic considerations do not play an important role. **Meanings** in such texts are, in practical terms, completely expressible in all relevant source and target languages.

✍ **6. Fully automated MT** is *not feasible* at present, but the main research direction is toward full automation.

✍ **7. SL ambiguity** resolution is the main technical goal to be achieved by MT systems.

✍**8.** Paradigmatic and other design considerations must crucially take into account the above requirement.

✍**9. Interlingual MT** systems tend to favor the meaning-based approach, while transfer systems tend to render meaning without the added requirement of representing it. Theoretically, meaning-oriented MT is not restricted to the **interlingua paradigm**. One can in principle incorporate meaning analysis into the transfer approach. However, in practice, as such attempts proliferate, it will become clear that the interlingua paradigm is more convenient for the support of the analysis of meaning. We also believe that the amount and complexity of knowledge acquisition for **interlingual MT** systems is at worst roughly equal to that which would have to be mastered for **meaning-oriented transfer MT**. At best, the acquisition component of an **interlingua approach** will be more compact and well-organized.

Chapter 24

Where Am I Coming From:

The Reversibility of Analysis and Generation in Natural Language Processing

Yorick Wilks

💣Main Notes of this article is available in Chapter 5 of the book "Wilks, Y. (2009). *Machine Translation: Its Scope and Limits*. Sheffield: Springer."

Chapter 25

The Place of Heuristics in the Fulcrum Approach to Machine Translation

Paul L. Garvin

✍**1.** The theoretical conception on which the **Fulcrum approach** is based is the definitional model of language. In this conception, the system of a language is considered to be, not a single hierarchy with a single set of levels ascending from phonology to semantics via syntax, but a multiple hierarchy structured in two dimensions, at least one of which in turn has three planes, with a separate set of levels proper to each of the planes.

✍**2.** Language is viewed as a system of signs structured in **two dimensions**, those of the grammar and the lexicon. These two dimensions differ in terms of the purpose to which the signaling means of the language are put:

⌘ the **lexical dimension** is defined as the system of reference to culturally recognized types of phenomena;

❄ the **grammatical dimension** is defined as the structure of discourse.

3. The **grammatical dimension** of language is characterized by three planes, each with its own set of distinctions: the plane of structuring, characterized in all languages by two levels of structuring—those of phonemics and morphemics; the plane of integration, characterized in all languages by several levels of integration (the number of which varies from language to language); the plane of organization, characterized in all languages by two organizing principles—those of selection and arrangement.

4. All of these distinctions are defined by **functional criteria**:

- The two **levels of structuring** differ in terms of the extent to which the units of each level participate in the sign function (meaning) of the language. The units of the phonemic level function primarily as differentiates of the sign function, the units of the morphemic level function as its carriers.
- The **levels of integration** differ in terms of the order of complexity of the units that constitute them: they range from the level of minimal units, which is the lowest, to the level of the maximal fused units, which is the highest. Fused units are considered to be not mere sequences of units of a lower order, but to function as entities of their own order, with certain overall qualities above and beyond the mere sum of their constituents. A correlate of the concept of fused units is the conception that the internal structure and the external functioning of a given unit are separate and potentially independent characteristics: units with the same internal structure may have different external functioning; units with different internal structure may have the same external functioning. Units with the same internal structure are called identically constituted; units with the same external functioning are called functionally equivalent.
- The two organizing principles on the plane of organization characterize different manners in which the signaling means of the language are employed: selection from an inventory versus arrangement in a sequence.

5. The three planes of the **grammatical dimension** of language are in a hierarchical relation to each other. The plane of structuring is defined by the most significant functional criterion and is therefore superordinate to the other two planes. Of the latter, the plane of integration is in turn superordinate to the plane of organization. Consequently, within each level of the plane of structuring a set of levels of integration can be defined, and within each level of integration of either level of structuring, the operation of both organizing principles can be discerned.

✍**6.** The **Fulcrum approach** differs from other approaches for automatic sentence structure determination primarily in the following respects:

- ◈ The Fulcrum approach favors a bipartite, rather than a tripartite, organization of the parsing system.
- ❁ The Fulcrum approach is characterized by two basic operational principles: (a) the concept of the fulcrum; (b) the pass method.
- ✯ The Fulcrum approach aims at producing a single interpretation of each individual sentence, rather than at producing all conceivable interpretations.

Chapter 26

Computer Aided Translation: A Business Viewpoint

John S. G. Elliston

✍**1.** The **demand** on our translation resource **grows** every year. This demand is related to our increasing product range, refinements to existing products and the normal **on-going need** to maintain existing documentation. An additional factor is the legal demands placed upon a multinational operation to translate to meet legal requirements. One obvious answer is to increase our resource to handle the growing load. Unfortunately, increasing the translation resource increases our **cost** base and makes us **less competitive**. The solution we need must be found in productivity, i.e., using the resources we already have, more efficiently.

PART III

SYSTEM DESIGN

Chapter 27

Three Levels of Linguistic Analysis in Machine Translation

Michael Zarechnak

✍**1.** This paper is one of the earliest explicit descriptions of MT design. Predating the ALPAC report by six years, it belies the often-stated view that linguistic and computational sophistication came to **MT** only after that damning report. ***Michael Zarechnak***'s presentation to the June 1958 meeting of the **ACM** describes the Georgetown system which was a forerunner of Systran, perhaps the single most successful **MT** system, and presents an approach which was to become entirely familiar over the next 30 years. In his general analysis technique (**GAT**), implemented on an **IBM 701** machine, we see one of the first examples of the separation of algorithms from the linguistic knowledge that they utilize, and in the three-level approach to linguistic analysis, we see the stratificational

approach that became so widespread. Zarechnak gives details of both his linguistic method, which he calls morphemic, syntagmatic and syntactic analysis, and of the data-structures used by the program: necessarily crude but actually not all that different from structures still in use some 25 years later (e.g., SUSY—Maas 1987).

Chapter 28

Automatic Translation—A Survey of Different Approaches

B. Vauquois

✍**1.** There was not much activity in MT in the late 1960s, and we jump almost 20 years, to COLING 1976 for our next landmark paper, Bernard ***Vauquois***' survey of different approaches. This is the article in which the distinction between first-generation and second-generation architectures is made, and is possibly the first appearance of the famous ''pyramid'' diagram that is almost obligatory in any general article about MT.

✍**2.** Notice that the **diagram** originally appeared with the apex at the bottom, facilitating the metaphors of surface and deep representations, which seem somehow less intuitive when the diagram is inverted.

✍**3.** The key elements of the **second generation** are all laid out here: the **modularity** and **stratification** of the translation process into **analysis** (parsing), **transfer**, and **generation**; the possible extrapolation of the analysis phase to an extent that transfer is unnecessary (the pivot or interlingua approach); the use of formalized and computable language models, and their nature (finite-state or context-free); and the separation of algorithms from the linguistic data.

✍**4.** The paper ends with the suggestion that a **third generation** of MT systems would make use of the results of research in artificial intelligence, incorporating richer semantics, and some knowledge of the real world.

✍**5.** ***Vauquois*** also describes ways in which the parallel goal of **human-assisted MT** could be achieved. Although not often cited, this paper clearly set the agenda and defined the vocabulary of MT system design for the following 20 years.

✍**6.** The **AI approach** mentioned by ***Vauquois*** is well illustrated by ***Yorick Wilks'*** description of his MT system, developed at Stanford University in the early 1970s. ***Wilks'*** system was designed both to translate and understand text, the latter to be demonstrated by an ability to answer questions (though ***Wilks*** was later to claim that translation was often as good a test of understanding as any, especially if it involved resolving ambiguities of word-sense, syntactic structure, pronouns and so on).

Wilks distances himself somewhat, in the opening paragraphs, from the formal logic approach to semantics that was prevalent at that time, and he places his approach firmly in the **interlingua** camp.

✍**7.** ***Wilks*** describes his approach in a lot of detail, which was somewhat unusual at the time, and the paper is above all interesting in that ***Wilks*** illustrates and discusses explicitly his proposed interlingual representations, and his examples tackle a variety of ambiguities and other difficulties.

Chapter 29
Multi-level Translation Aids
Alan K. Melby

✍**1.** While Wilks and others explored the possibilities of MT systems influenced by advances in AI, ***Alan Melby*** took on the proposal to develop systems where the computer would cooperate with a human user to produce high-quality translations. Although this approach had been suggested by various commentators, notably of course the **ALPAC report**, but also Lippmann (1971) and Kay (1980), it was Melby who can be credited with having done the most to see these ideas realized. In a series of articles developing the theme, and in software which was eventually marketed commercially (in the form of the ALPS system), ***Melby's*** ''Interactive Translation System'' (ITS) became a blueprint for the Translator's Workstations that are now more or less familiar.

✍**2.** ***Melby's*** key idea was that the computer should be flexible enough to offer aid to the translator at different levels ranging from simple text processing to terminology aids to full machine translation. ***Melby's*** thoughtful analysis of the role of the translator in MT and his personal experience of this job have had an important impact on the field.

Chapter 30
EUROTRA: Computational Techniques
Rod Johnson, Maghi King, and Louis des Tombe

✍**1.** The European Commission's **EUROTRA MT** project was, and perhaps always will be, the largest **MT** project ever undertaken, both in terms of cost and personnel. It is not controversial to say that its outcome was a huge disappointment, and this is not the place to discuss that aspect of it. In its early days, the project was shrouded in a veil of secrecy, imposed by the funders, so that few details of its design were published, beyond fairly banal and superficial descriptions of the impact on the system design of the organizational structure of the project (in particular, the

geographical dispersal of those working on the project, and the desire to accommodate diverse scientific predilections).

✍**2.** Reproduced here is an extract from the article which appeared in the 1985 special issue of Computational Linguistics, containing descriptions of more or less all the important MT systems at that time. The article was mostly about the general design and organizational structure, but the section reproduced here also shows that the project resulted in some innovative ideas about some computational aspects of **MT** system design. The extract discusses the problems of finding the appropriate level of specificity and generality for a linguistic formalism and implementing it in a distributed and robust fashion.

✍**3.** The discussion illustrates the underlying tensions between **procedural** and **declarative** programming styles, providing a framework that was comfortable for linguists with varying experience of computational linguistics, the result needing also to be efficient and reliable. Although Johnson, King and *des Tombe* rejected the use of an existing programming language, perhaps extended by a library of purpose-built macros, subroutines or functions, eventually this was the approach adopted for the **EUROTRA** system, though it should be said that in the choice of Prolog for this task, many of the concerns and ideas expressed in this early article were influential.

Chapter 31

A Framework of a Mechanical Translation between Japanese and English by Analogy Principle

Makoto Nagao

✍**1.** ***Makoto Nagao*** has been one of the most influential and important names in MT research, not only in Japan but worldwide. The paper he delivered at a minor symposium in France in 1981, published three years later in a little-read collection, languished in obscurity until the start of the next decade, when suddenly and unexpectedly a whole new paradigm for MT emerged.

✍**2.** ***Nagao's*** paper is inevitably cited as the first one in which **Example-based MT** is proposed, although actually ***Nagao*** does not use this term, but rather talks of ''machine translation by example-guided inference,'' or ''machine translation by the analogy principle.'' The main features of **EBMT** are there nevertheless: the use of examples rather than rules to establish the correspondences; and the need for some means to quantify the similarity between the input and the various examples (Nagao assumes the use of a thesaurus).

Chapter 32
A Statistical Approach to Machine Translation

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. La.erty, Robert L. Mercer, and Paul S. Roossin

✍ **1.** Apparently quite independently of Nagao, the **BSO** research group in Utrecht, and in particular Victor Sadler, had a number of ideas about using a small corpus of examples as a general-purpose knowledge source for **NLP** purposes. In including this paper in our collection, we are perhaps departing slightly from our goal of including influential and much-cited papers, since this one, presented at a semi-private (invitation-only) seminar, is probably not widely known. But we include it because it contains several ideas which were later to become widespread, and thus Sadler should be acknowledged as one of the first researchers to suggest them. For example, since the sentences in the corpus were stored as grammatically annotated tree structures, this is an early example of a tree bank. Sadler goes into extensive detail about how such a resource can be developed and used, using the term ‘‘example-based’’ explicitly, and probably predating the use by various Japanese researchers of that term.

✍ **2.** Interestingly, the seminar where this paper was presented was organized by **ATR**, one of the groups which is closely associated with this approach. The **BSO** group had already presented their idea of a bilingual knowledge bank, another analogical technique especially useful for word-sense disambiguation, at **COLING** in 1990. In fact, the **BSO** group never really got the opportunity to explore their ideas about **EBMT** fully, being victims of changed funding priorities in the mid-1990s.

✍ **3.** Another new technique which emerged at the beginning of the 1990s was the ‘‘**statistical**’’ approach, with the **IBM** group led by Peter Brown in the forefront. The paper reproduced here appeared in Computational Linguistics and gives the most complete description of their early experiments, which had been presented at various conferences in the two preceding years, the first presentation to an **MT** audience having been at the **TMI** conference at Carnegie Mellon University, Pittsburgh, in 1988. In this article are the essential elements of the approach: a later article (Brown et al. 1993) gives more details about the mathematical models, and indeed the statistical approach itself was later modified to take more account of linguistic generalizations, e.g., morphology, before the group split up some six or seven years later. At the time, the statistical approach, along with **EBMT**, was seen (by some) as a serious challenge to the by now traditional rule-based approach, this challenge typified by

the (partly engineered) confrontational atmosphere at **TMI-92** in Montreal.

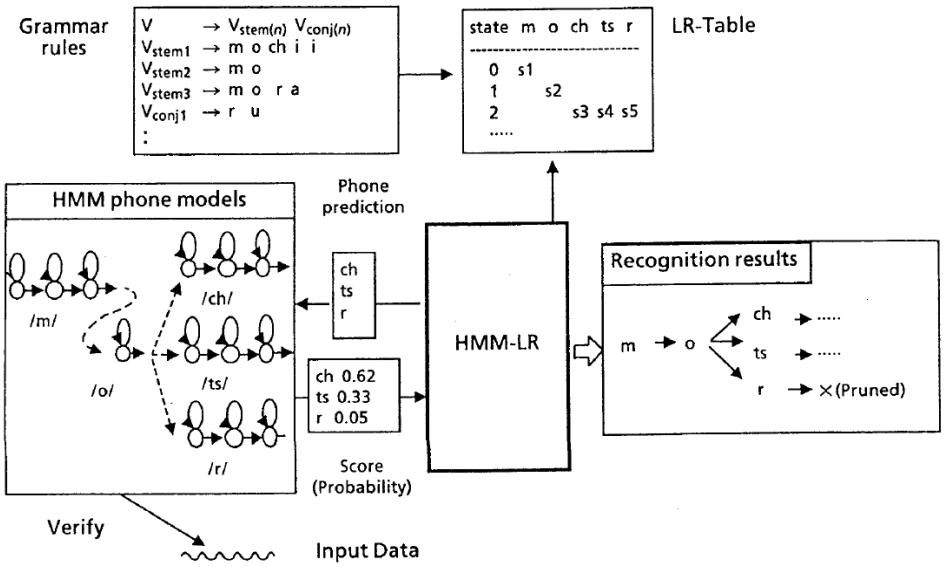
✍ **4.** Although some researchers are still following a strictly empiricist approach, the more significant outcome is now a number of hybrid system designs involving statistical, **corpus-based** and **rule-based** processes. The related activity, not strictly **MT** but somewhat relevant, of bilingual corpus alignment has enjoyed a great deal of attention in recent years, and has contributed to the development of a number of useful tools for translators.

Chapter 33

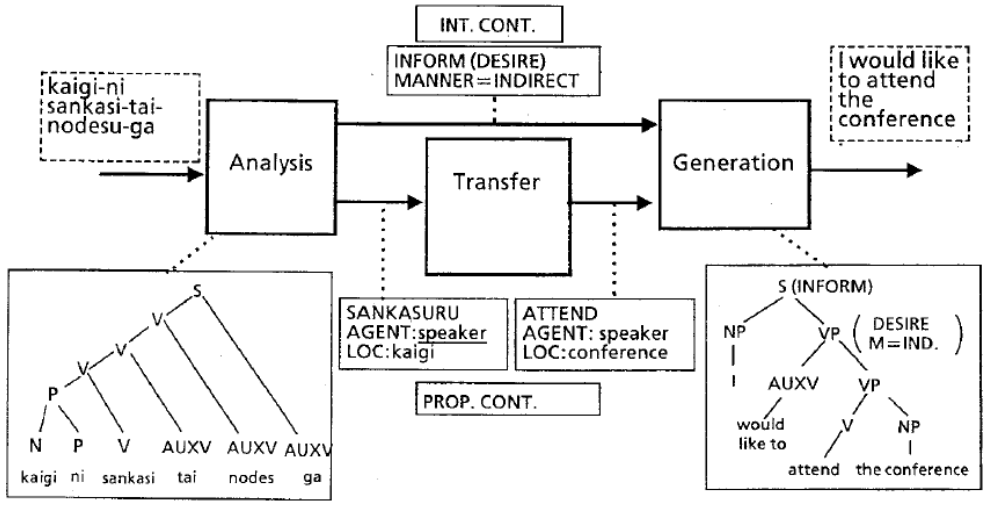
Automatic Speech Translation at ATR

Tsuyoshi Morimoto and Akira Kurematsu

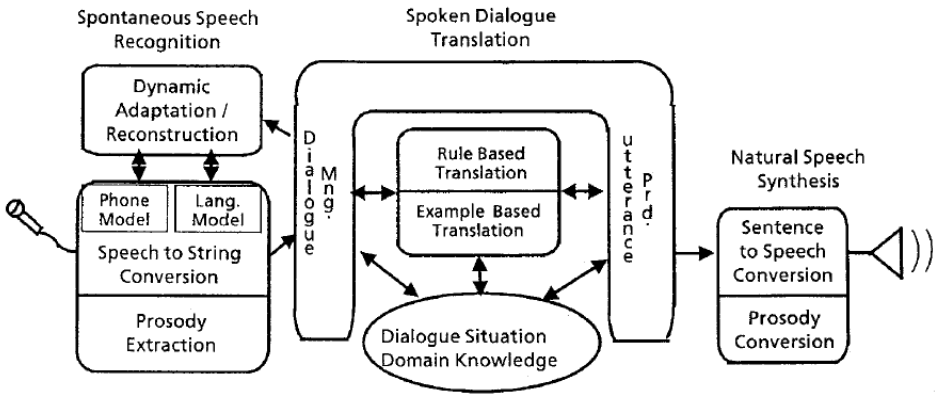
✍ **1.** Basically, two kinds of models are necessary for speech recognition: a phonetic model and a language model. For phonetic modeling, a Hidden Markov Model (**HMM**) approach was employed. A phone is apt to be acoustically affected by preceding and/or succeeding phones, so hundreds of allophone models are generated automatically from a huge speech database by use of the ‘‘successively-state-splitting’’ (**SSS**) algorithm (Takami and Sagayama 1992). For the **language model**, a general context free grammar (**CFG**) was used. Compared to other conventional language models such as bigram or trigram, it is superior in extendability and maintainability. A new mechanism, a predictive **LR** parsing mechanism which is an extension of the generalized **LR** parsing algorithm, combines these two models dynamically and recognizes input continuous speech (Kita et al. 1989). In this method, **CFG rules** are compiled and converted to an LR table. The parser refers to the table and predicts the next possible phones, then verifies their existence in the input speech by comparison with corresponding HMMs (figure 33.1):



2. The style of **spoken sentences** is, especially in Japanese, quite different from that of written sentences. Spoken sentences include various intentional expressions or ellipses. To treat such sentences, a new method called the ‘‘intention translation method’’ (Kurematsu et al. 1991) was developed (figure 33.2). An input utterance is analyzed by the analyzer based on the HPSG (and its Japanese version JPSG) grammar formalism and unification operation. In each lexical entry, syntactic, semantic and even pragmatic constraints are defined in the form of feature structures. In this paradigm, the inefficiency caused by the unification operation is the biggest issue, and various efforts have been made such as introducing medium-grained CFG rules (Nagata 1992) or implementing a quasi-destructive graph unification algorithm (Tomabechi 1992) to solve this issue. With these efforts, the processing time has been drastically decreased. The next transfer component is composed of three phases: zero-anaphora resolution, illocutionary force type determination, and conversion of source-language semantics to target-language semantics.



3. Especially in **spontaneous speech translation**, the integrated control of speech and language processing becomes very important. Appropriate information necessary for language models should be provided to speech recognition from the language processing side, and speech information such as prosody should be provided to language processing from the speech processing side as well. **At the same time**, the status of the dialogue should be recognized and maintained properly. Such situational information would be about the environment (such as the domain or the subject of the dialogue), the participants' status (such as their intentional or mental states) and the dialogue progression status (such as the topic or the focus). Such information would be referred to by both the speech processing and the language processing. The overall image of the future system would be like *figure 33.6*:

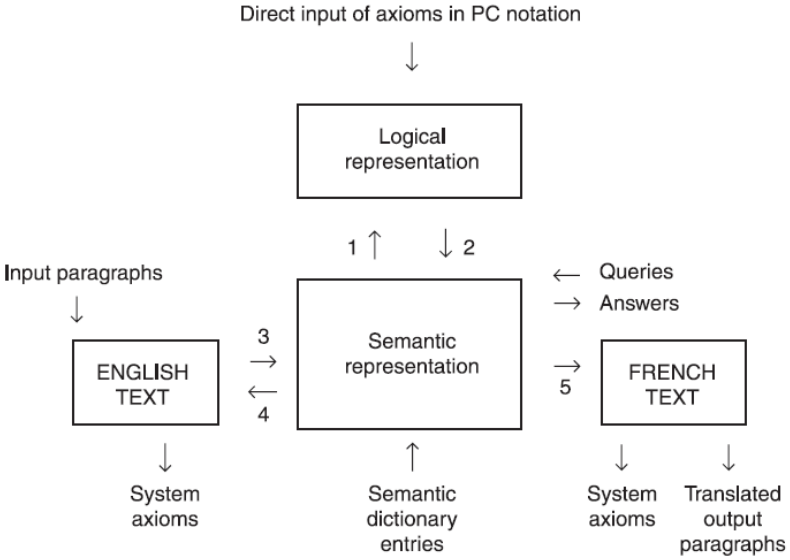


Chapter 34

The Stanford Machine Translation Project

Yorick Wilks

✍**1.** The diagram in *figure 34.1* represents the overall structure of the system under construction. I assume in what follows that processes 2, 4, and 5 are the relatively easy tasks—in that they involve throwing away information—while 1 and 3 are the harder tasks in that they involve making information explicit with the aid of dictionaries and rules.



Chapter 35

The Textual Knowledge Bank: Design, Construction, Applications

Victor Sadler

✍**1.** Basically, the **TKB** concept is simple enough. It represents a way of storing full text, not as an extended string of characters, but as a grammatically and referentially coded tree structure in which the nodes are linguistic objects on various levels and from which the original character string can be reconstructed at any level (***figure 35.1***).

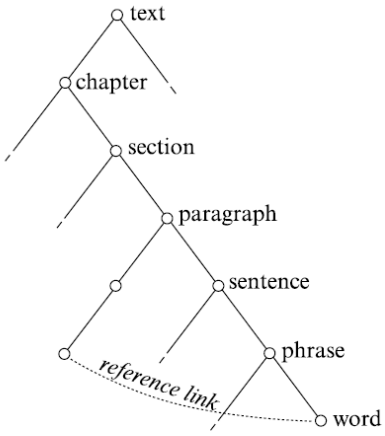


Figure 35.1
Conceptual view of a Textual Knowledge Bank.

✍ **2.** The aim is to make the **knowledge** contained in ordinary texts accessible to the computer—without formalizing the linguistic knowledge into rules and without building an abstract knowledge representation divorced from the linguistic level. To this end, the text has to be structured: first by identifying its components (words, morphemes or whatever); second by drawing syntactic relations between those components (dependency parsing); and third by drawing reference relations between components which in one way or another refer to the same thing (anaphora, etc.). In this way, it was argued, both the linguistic and the non-linguistic knowledge (knowledge of the world) required for natural language processing could be combined into a single knowledge source.

✍ **3.** ***Figure 35.2*** illustrates the **TKB structure**, comprising both syntactic and referential links, for the following pair of sentences: (1) Use the Delete option to delete individual documents. The owner and other sharers cannot access those documents.

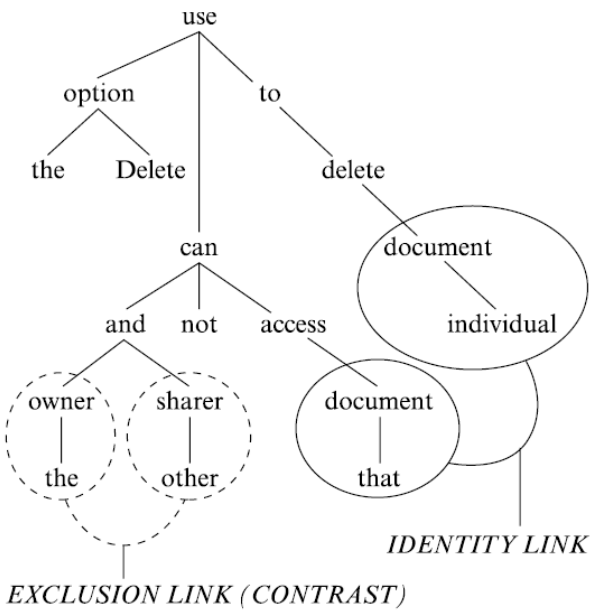


Figure 35.2
TKB structure for example (1).

4. Where bilingual or multilingual applications are concerned, a further dimension is added to the structure. For MT purposes, for instance, parallel texts (translations) in different languages are first structured in the way described above, and then additional, bilingual relations are drawn between equivalent units in the two parallel structures (***figure 35.3***). The result is termed a Bilingual Knowledge Bank (or **BKB**).

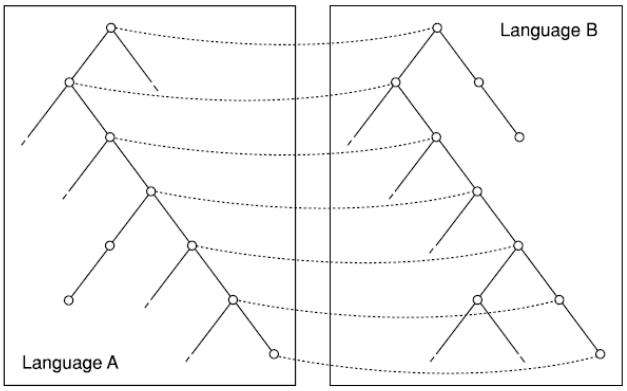


Figure 35.3
Coupling of two equivalent TKBs.

5. ***Figure 35.4*** illustrates by means of an example sentence the coupling of two (monolingual) TKBs into a (bilingual) BKB. In this figure, the

dependency structure of the English and French sentences is shown in a different graphical form from that of ***figure 35.2***, with (sub)trees defined by (boxes within) boxes. It will be clear that a BKB can function as a kind of bilingual (and bidirectional) dictionary, with an abundance of contextual examples. As far as linguistic knowledge is concerned, the motivation behind the construction of a large database such as a TKB is the conviction that rule-based systems have proved inadequate for NLP purposes, and that analogical or ‘‘**example-based**’’ or ‘‘**memory-based**’’ techniques are needed instead–or possibly in addition (the hybrid approach).

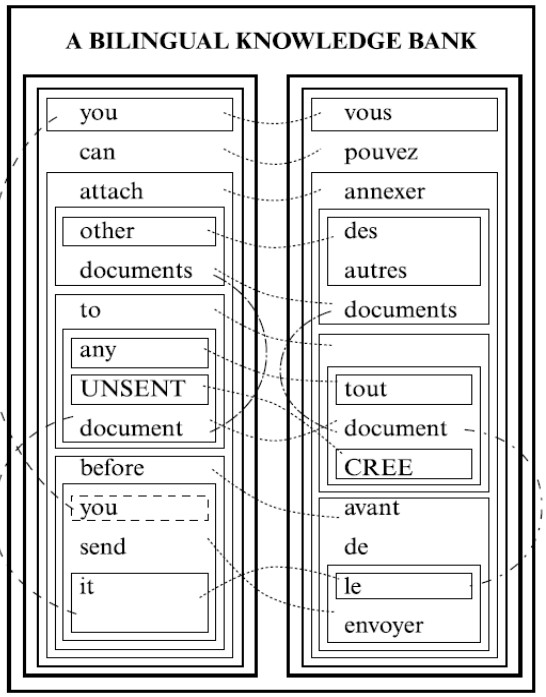


Figure 35.4
Fragment of a BKB.

6. The basic function of a **Textual Knowledge Bank** is to serve processes which attempt to match words and phrases from their input with the pre-analyzed examples in the **TKB**, in order to interpret their meaning and/or to check for consistency. This function suggests that **TKB** technology can form the basis for a wide variety of applications such as intelligent spelling and grammar checkers, extraction of technical terms, intelligent information retrieval and, in the longer term, machine translation.

Chapter 36

Machine Translation Without a Source Text

Harold L. Somers, Jun-ichi Tsujii, and Danny Jones

✍ **1.** ***Somers et al.*** actually introduce a second theme which has proven to be a predominant one in the 1990s, and which they dubbed ‘‘*translation without a source text*.’’ Adapting MT for users other than translators, and who may even be monolingual, multilingual generation of target text on the basis of negotiation with the user is presented. Subsequent system designs proposed variants which could be described as ‘‘**multilingual summarization**,’’ where the source data, which may or may not be in a textual form, is analyzed and represented to the user in a variety of textual forms which are not necessarily based on that of the original.

✍ **2.** It is important to emphasize that there is a basic difference between **Dialogue Machine Translation** (DMT) systems on the one hand and conventional MT systems on the other, namely the difference of user types. In **DMT**, users are dialogue participants who actually have their respective communicative goals and who really know what they want to say. On the other hand, the users of conventional MT are typically translators who, though they have enough knowledge about both languages, lack ‘‘**complete understanding**’’ of texts to be translated.

✍ **3.** In **DMT**, the system can ask in theory any questions to elicit the information necessary for translation which is not explicitly expressed in the ‘‘source text’’. This is impossible in conventional MT, because the users do not have ‘‘**complete understanding**’’ of the context in which the texts are prepared, and the users (who are translators) simply could not answer such questions. *(It is often the case that even human translators would like to consult the authors of the original texts in order to produce a good translation.)*

✍ **4.** In order to exploit this advantage in **DMT** however, we have to overcome several related difficulties. In **DMT** there are several different types of dialogues, any of which may start up or be resolved at any given time: these dialogues include

- ☸ user–user object-level dialogues
- ▣ user–user meta-level dialogues (e.g. in which one participant in the dialogue asks the other participant questions to clarify the meaning or intentions of his/ her statements)
- ⌗ user–system dialogues typically initiated by the system, concerning the progress of the object-level dialogue, disambiguating ambiguous object-level dialogue, i.e., what the user wants to say next
- ✠ user–system meta-level dialogues typically initiated by the user, concerning clarification of the object-level dialogue, i.e., what was just said

✍**5.** One of the foreseeable **difficulties** in **DMT** is how to distinguish these different modes of dialogue, that is, how systems can distinguish, first of all, utterances of types (a) and (b) to be translated and transmitted, from utterances of type (d) which should not be translated. In particular, dialogues of types (b) and (d) may be difficult in some cases, because the user posing questions of clarification cannot generally recognize whether the difficulties of understanding come from ‘‘**errors**’’ in translation or from the other participants’ utterances themselves.

Short Answer Items & Tests

4.2 Short Answer Items

1. Understanding the ‘‘linguistic’’ meaning of a text does not ……. the ability to process a text correctly: ‘‘linguistic’’ meaning and ‘‘situational’’ content are quite different things not always ……. by a unique (one-to-one) correspondence.

2. The right translation is possible only if the ……. situation is rightly understood.

3. In the transfer approach, there is the same number of analyzers and generators (as is in the pivot approach), but ……. transfers must be added.

4. Translation is rendering a set of meanings realized in a ……. language using the realization means of a ……. language.

5. According to Paul L. Garvin, the grammatical dimension of language is characterized by three planes: the plane of structuring; the plane of …….; the plane of ……. .

4.3 Answers

1) guarantee, linked	**2)** extralinguistic
3) n(n-1)	**4)** source, target
5) integration, organization	

☙☆❧ 4.4 Tests ☙☆❧

✍ Select the best choice.

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

a) writing, perception
b) speaking, recognition
c) communicating, detection
d) cognition, synthesis

2. The pivot approach seems best suited to the construction of multilingual M(A)T systems, for obvious reasons of ……. .

a) economy
b) minimality
c) both a and b
d) neither a nor b

3. Translation is a relation between texts in the source and target languages, such that the invariant between them is ……. .

a) meaning
b) form
c) structures
d) lexemes

4. MT deals with ……. texts, where the artistic considerations ……. an important role. Meanings in such texts are, in practical terms, completely ……. in all relevant source and target languages.

a) expository, do not play, expressible
b) expository, plays, expressible
c) expressive, plays, transferable
d) expressive, do not play, conveyable

5. Fully automated MT is ……. at present. The main research direction ……. toward full automation.

a) not realistic, is not
b) feasible, is not
c) practical, is
d) not feasible, is

☙✷❧ 4.5 Answer key ☙✷❧

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

Appendix

Slide ❶

Jaime Carbonell (2007)

HISTORY OF MACHINE TRANSLATION

1. Origins of MT: *Early* "*Successes*"

- 1933 – Smirnov-Troyanskii Patent for a word translation & printing machine
- 1939-1941 – Troyanskii added memory (first Russian computer)
- 1946 – MT as code-braking (ENIAC in US), Weaver et al
- 1946-1947 – Weaver, Booth, Weiner … Weaver realizes complexity
- 1949 – Weaver Memorandum (what it would take for MT)
- 1951 – Bar Hillel survey => Human/machine is best
- 1952 – MIT Conference on MT (first small scale E-F, F-E mostly)
- 1954 – Mechanical Translation Journal (Yngve)
- 1954 – Georgetown-IBM Experiment (50 sentences R-E) => massive US funding
- 1956-1962 – Massive MT efforts at U of Washington, IBM, Georgetown, MIT, Harvard, Oakridge, Rand, using any and all hardware including Mark II, ILIAC, …
- 1960-1964 – Kuno (Harvard) and Oettinger (Georgetown) parser
- 1955-1967 – UK active in MT (Booth, Cambridge group)
- 1956-1965 – MT in Japan starts (Wada at ETL, Fukuoka at Kyushu, …)
- 1960's => on – GETA in Grenoble (Vauquois)

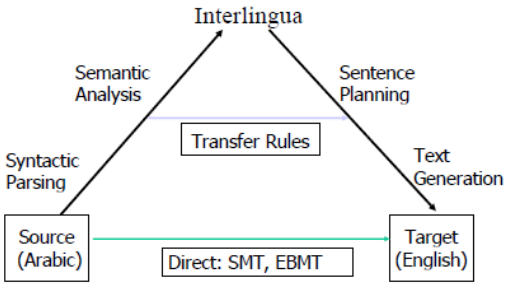
2. Origins of MT: *End of Optimism*

- 1960 – Bar-Hillel report and the FAHQT Myth
- 1964, April – ALPAC Report

3. The MIT Early History: Bar-Hillel

- Philosopher & Mathematician, but turned Linguist & MT booster
- First-ever full-time MT researcher (MIT: 1951-1953)
- Recognized lexical ambiguity as largest challenge for MT
- Identified other MT challenges

4. Types of Machine Translation:



5. The MIT Early History: *Victor Yngve*

- High-Energy Physicist turned Linguist
- 2nd-ever full-time MT researcher (MIT: 1953-1961)
- Word-for-word MT => syntax matters (for resolving homonyms e.g. "block" and for word-order inversion)
- Recognized phrasal lexicon
- Invented analysis-transfer-generation method
- Invented COMIT (operational grammar encoding)
- Implemented Chomsky's TG in COMIT (which proved a dismal failure for analysis)

6. The Georgetown Early History: *Leon Dosert*

- Linguist & Interpreter during WWII
- Attracted most MT funding (military)
- Focused on Russian => English
- Strongest advocate for MT research

7. The Georgetown Early History: *First "large-scale" MT*

- About 100,000-word Russian Text MTed in demo adding out-of-dictionary words (1958)
- System scaled further in next 5 years
- GAT (Georgetown Automated Translator) => Well-known SYSTRAN in later years

8. The ALPAC Report: *Members*

- Pierce (Chair) Bell Labs
- Several discouraged MT researchers (Oettinger, Hays)
- Linguists (Hamp, Hockett)
- Token Computer Scientist (Alan Perlis from Carnegie Tech)

9. The ALPAC Report: *Findings*

- Myth – MT does not and cannot work
- Reality – MT is more difficult than originally envisioned
- Reality – Basic Research in NLP should be done before doing MT
- Reality – MT is too expensive (computers cost more than people)

10. The ALPAC Report: *Net Effect*

- The end of Government-funded MT research in US for 10+ years
- Continuation of private MT (e.g. Systran, Logos) in US
- Not much effect on Japan or France (efforts continued)
- USSR and UK followed US example, it appears

11. MT: (1967-1985) ALPAC Myth Fades Away in US

- SYSTRAN quite successful in E-R (Air Force at Wright-Patterson etc.)
- Partial success E-S, E-F, E-G (SYSTRAN, Logos, Weidner)
- SYSTRAN => use in Europe (later by EC)
- Knowledge-Based MT (KBMT) concept advanced (Carbonell, Nirenburg, …)
- "Underground MT" in US Universities dares to seek funding again
- Machine-aided Translation (MAT) concept advanced (Kay, …)
- Very-narrow-domain MT demonstrated (Kittredge et al, METEO)

12. MT: (1975-1985) Golden-Age of MT in Japan:1980's

- Nagao proposes Example-Based MT (not taken seriously then)
- Nagao proposes Transfer-Based MT for E-J (Mu project)
- Mu's success triggers MT-mania in giant Japanese companies, e.g., ATLAS in Fujitsu, PIVOT in NEC, HICATS in Hitachi, …
- Japanese MT Research budgets soar, US and Europe take note
- JEIDA Report paints upbeat future for MT

13. MT: (1975-1985) MT in Europe, not as Rosy

- "Interlingua" approach tried (ROSETTA, DLT)
- First language-neutral Interlingua (Yale-MT, Carbonell & Cullingford 1979, 1981)
- Eurotra proposed and started to build ultimate collaborative MT system, but later tanks due to incompatible transfer paradigms
- …but SYSTRAN adopted by EC for volume internal translations

14. MT Matures (1985-1995): MT Spring in US

- Center for Machine Translation at CMU opens in 1986
- Interlingual KBMT success at CMU for domain-oriented MT (KANT) with controlled-language input, but did not generalize to open-ended and uncontrolled domains (PANGLOSS)
- Resurgence of statistical corpus MT at IBM (Brown et al), which also succeeds for E-F but needs huge training corpus
- Speech-to-Speech MT launched at CMU (first JANUS, the DIPLOMAT)
- CSTAR launched (International consortium for speech-speech MT)
- SYSTRAN, LOGOS, GLOBAL-LINK (formerly Weidner), … survive
- Conferences: MT-Summit, TMI, … (MT regains respectability)

15. MT Matures (1985-1995): MT Summer and Fall in Japan

- Japanese systems reach performance plateau, typical for transfer-MT

- Funding reduced, especially when economic difficulties intrude
- MT useful with extensive post-editing (e.g. ATLAS-II MT bureau)
- ATR Successful in speech-speech MT for limited domains
- Example-based MT re-emerges (Iida at ATR, Nagao at Kyoto)

16. MT Matures (1985-1995): MT Mostly Sub-Rosa in Europe

- EUROTRA a massively distributed uncollaborative failure
- Companies abandon MT efforts (DLT, Rosetta, Metal)
- SYSTRAN in large-scale deployment and use in EU shines through
- Vermobil speech-speech MT in Germany concluded with reasonable large-scale success for speech-MT

17. The Modern Period: MT post 1995 Technological Trends

- Transfer MT works with high development & post editing costs
- Interlingual KBMT works well in technical domains (but requires high development cost)
- Speech-to-Speech MT increasing in popularity, but not yet robust
- Example-Based MT => Generalized EBMT
- New-wave of Statistical MT (CMU, ISI, JHU)
- Example-Based MT (Kyoto U, CMU)
- MT research ongoing and respectable, but with modest funding (in US, Japan, and Europe)
- Rapid-development MT becomes hot topic (US Govt., CMU, NMSU, internet)

18. The Modern Period: MT post 1995 Application Trends

- SYSTRAN, LOGOS, L&H, IBM, Fujitsu, remain steady MT suppliers
- Interlingual KBMT in first massive use (at Caterpillar)
- PC-based MT Systems explode (Fujitsu, IBM, Globalink, L&H)

19. The Modern Period: 1995-Present

- Internet MT off to a good start (Babblefish, Google)
- Translingual IR + MT hot (CMU, IBM, Google, …)
- Speech-speech MT reinvigorated
- New DARPA MT initiative
  - *Statistical MT dominates*
  - *Evaluation centric (NIST, BLEU, ...)*
  - *Focus on non-European languages (Arabic, Chinese)*
- Japan & Europe => MT slidelines
- India, China, Russia become serious MT players

20. MT: Present & Future Trends

- Evaluation is here to stay
  - *New, better methods (e.g. METEOR at CMU)*
- New paradigms for MT flourish
  - *Transfer-rule learning (CMU)*

- *– CMBT = EBMT without parallel text (Meaningful M.)*
- *– Hybrid methods EBMT/SMT/RuleMT*
- *– Multi-Engine MT*

• Biggest challenge: Breaking the Accuracy Bottleneck
- *– MT with accuracy comparable to Human Translators*
- *– Huge translation market (20+ billion/year)*

21. Lessons from MT History:

• Translation ≠ Transduction
• MT is a paradigm task for NLP
• Context, context, context
- *– word-for-word*
- *– transfer grammars + lexical substitution*
- *– KBMT with semantic interpretation rules*
- *– statistical MT with bi-grams & trigrams*
- *– phrases (bigger n-grams) matter (EBMT, SMT)*
- *– new methods are based on yet longer n-grams*

• Machine learning enters MT, more and more
• In MT perseverance and longevity matter

Slide ❷

Alon Lavie (2007)

Transfer Methods for Machine Translation

1. Direct Approaches:

⌘ No intermediate representation stages in the translation
✠ First MT systems developed in the 1950's-60's (assembly code programs)
- *– Morphology, bi-lingual dictionary lookup, local reordering rules*
- *– "Word-for-word, with some local word-order adjustments"*

▣ Modern Approaches: EBMT and SMT

2. Analysis and Generation Main Steps:

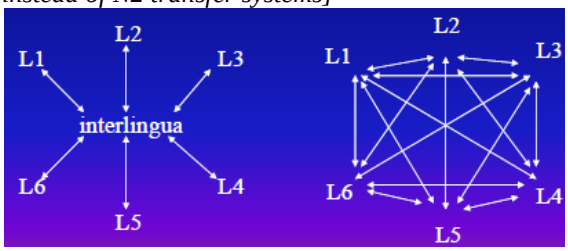
☆ ***Analysis***:
- *– Morphological analysis (word-level) and POS tagging*
- *– Syntactic analysis and disambiguation (produce syntactic parse-tree)*
- *– Semantic analysis and disambiguation (produce symbolic frames or logical form representation)*
- *– Map to language-independent Interlingua*

✵ ***Generation***:
- *– Generate semantic representation in TL*
- *– Sentence Planning: generate syntactic structure and lexical selections for concepts*
- *– Surface-form realization: generate correct forms of words*

3. Transfer Approaches:

- ***Syntactic Transfer***:
  - *Analyze SL input sentence to its syntactic structure (parse tree)*
  - *Transfer SL parse-tree to TL parse-tree (various formalisms for specifying mappings)*
  - *Generate TL sentence from the TL parse-tree*
- ***Semantic Transfer***:
  - *Analyze SL input to a language-specific semantic representation (i.e., Case Frames, Logical Form)*
  - *Transfer SL semantic representation to TL semantic representation*
  - *Generate syntactic structure and then surface sentence in the TL*

4. Interlingua versus Transfer: *[With interlingua, need only N parsers/ generators instead of N2 transfer systems]*



5. Transfer Approach:

- Language-dependent intermediate representations
- Disadvantage: costly as number of languages grows
- n (n - 1) transfer components

6. Advantages of Interlingua:

- Add a new language easily *[get all-ways translation to all previous languages by adding one module for analysis and one module for generation]*
- Mono-lingual development teams
- Paraphrase *[Generate a new source language sentence from the interlingua so that the user can confirm the meaning]*

7. Disadvantages of Interlingua:

- "Meaning" is arbitrarily deep. *[What level of detail do you stop at?]*
- If it is too simple, meaning will be lost in translation.
- If it is too complex, analysis and generation will be too difficult.
- Should be applicable to all languages *[How do we ensure that?]*
- Human development time.

8. Transfer Approaches: *[Main Advantages and Disadvantages]*

- ***Syntactic Transfer:***
  - No need for semantic analysis and generation
  - Syntactic structures are general, not domain specific => Less domain dependent, can handle open domains
  - Requires word translation lexicon

- ***Semantic Transfer:***
  - Requires deeper analysis and generation, symbolic representation of concepts and predicates => difficult to construct for open or unlimited domains
  - Can better handle non-compositional meaning structures => can be more accurate
  - No word translation lexicon – generate in TL from symbolic concepts

9. Major **Sources** of **Translation Problems** and Divergences:

- *Lexical Differences*: [Multiple possible translations for SL word, or difficulties expressing SL word meaning in a single TL word]
- *Structural Differences*: [Syntax of SL is different than syntax of the TL: word order, sentence and constituent structure]
- *Differences in Mappings of Syntax to Semantics*: [Meaning in TL is conveyed using a different syntactic structure than in the SL]
- *Idioms and Constructions*

10. MT Handling of Lexical Differences

- Direct MT and Syntactic Transfer:
  - Lexical Transfer stage uses bilingual lexicon
  - SL word can have multiple translation entries, possibly augmented with disambiguation features or probabilities
  - Lexical Transfer can involve use of limited context (on SL side, TL side, or both)
  - Lexical Gaps can partly be addressed via phrasal lexicons
- Semantic Transfer:
  - Ambiguity of SL word must be resolved during analysis => correct symbolic representation at semantic level
  - TL Generation must select appropriate word or structure for correctly conveying the concept in TL

11. MT Handling of Structural Differences

- Direct MT Approaches:
  - No explicit treatment: Phrasal Lexicons and sentence level matches or templates
- Syntactic Transfer:
  - Structural Transfer Grammars
- Trigger rule by matching against syntactic structure on SL side
- Rule specifies how to reorder and re-structure the syntactic constituents to reflect syntax of TL side
- Semantic Transfer:
  - SL Semantic Representation abstracts away from SL syntax to functional roles => done during analysis
  - TL Generation maps semantic structures to correct TL syntax

12. MT Handling of Syntax-to-Semantics Differences

- Direct MT Approaches:

– No Explicit treatment: phrasal lexicons and sentence level matches or templates

❄ Syntactic Transfer:

– "Lexicalized" Structural Transfer Grammars

❀ Trigger rule by matching against "lexicalized" syntactic structure on SL side: lexical and functional features

⌘ Rule specifies how to reorder and re-structure the syntactic constituents to reflect syntax of TL side

✠ Semantic Transfer:

– SL Semantic Representation abstracts away from SL syntax to functional roles => done during analysis

– TL Generation maps semantic structures to correct TL syntax

13. MT Handling of Constructions and Idioms

▣ Direct MT Approaches:

– No Explicit treatment: Phrasal Lexicons and sentence level matches or templates

☆ Syntactic Transfer:

– No effective treatment

– "Highly Lexicalized" Structural Transfer rules can handle some constructions

❋ Trigger rule by matching against entire construction, including structure on SL side

⌖ Rule specifies how to generate the correct construction on the TL side

◈ Semantic Transfer:

– Analysis must capture non-compositional representation of the idiom or construction => specialized rules

– TL Generation maps construction semantic structures to correct TL syntax and lexical words

14. Transfer-based MT Systems:

⁂ Primarily Syntactic-transfer, based on large manually developed transfer grammars

※ Most notable systems:

– SYSTRAN translation engines

– PAHO system (Spanish/English)

– EUROTRA

– VERBMOBIL

🖹 Main Issues:

– Large volume and complexity of transfer grammars

– Interaction between "general" and "exception" rules

– Interaction between transfer grammar and lexicon

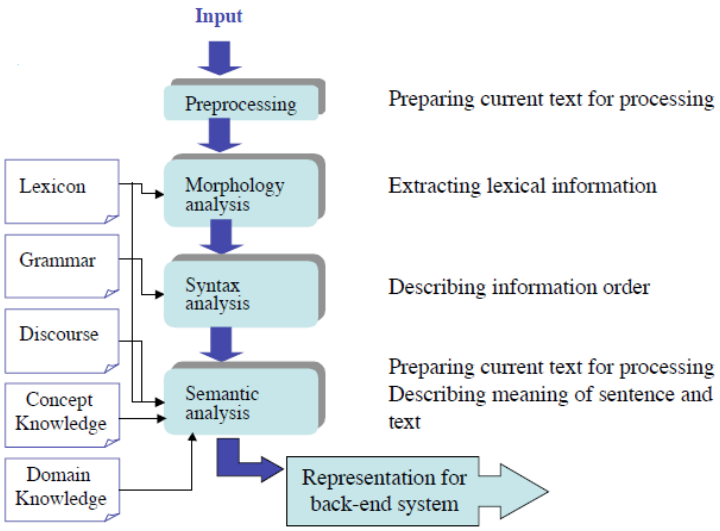
Slide ❸

***Cristina Vertan** (2004, University of Hamburg)*

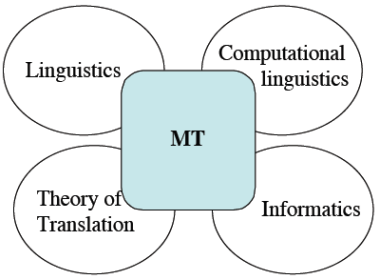
PART I

Introduction to Machine Translation

1. NLP Standard Architecture:



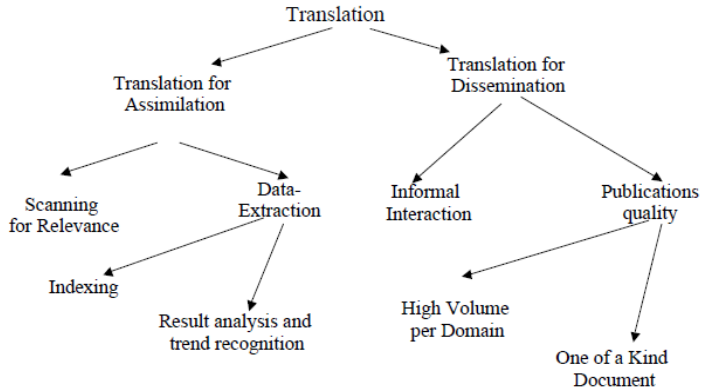
2. MT is **not a discipline** by itself, but an application of several disciplines:



3. Features which we expect (*at least*) from MT systems:

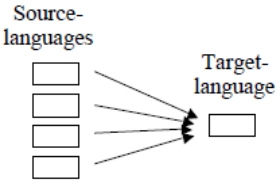
- ※ Semantic adequacy
- ☆ Stylistic and pragmatic adequacy
- ❄ Cultural adequacy
- ❋ Consistency inside a text and between texts
- ⌖ Reduced costs compared to human translations
- ❁ High speed

4. Functional Typology of MT-Systems:



5. Translation for Assimilation/Dissemination:

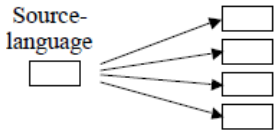
Assimilation:



Process:

- Each language
- Each style level
- almost each theme
- All-purpose translation
- Few Semantics
- Requires Post-editing

Dissemination:



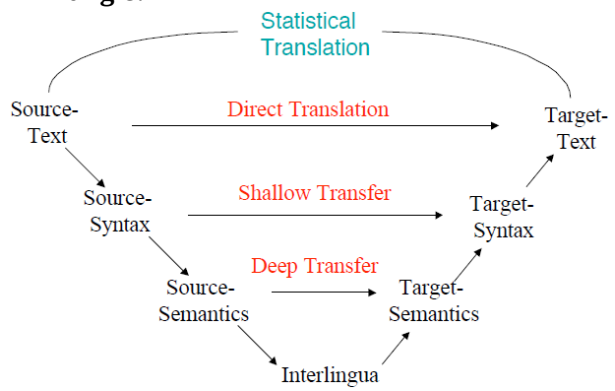
Process:

- One source language
- Style is defined
- A theme or a domain
- Special translations
- Full semantic analysis
- No Post-editing

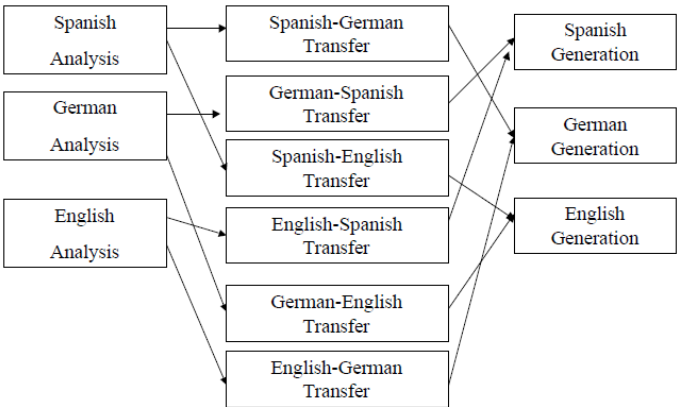
6. History:

Year	U.S.A	Europe	Japan
1950ies	Start of big MT-systems		Early MT-Research
1960ies	ALPAC End of MT	Start of MT	
1970ies 1980ies	(SYSTRAN, METAL) NLP Basic Reserach	GETA EUROTRA	
1990ies	Newr Start in MT-Research (SYSTRAN)	EUROTRA (METAL SYSTRAN)	MT-System MT Boom in Industry MT-Products
	Official MT-research (SYSTRAN) Multilingual Systems	End of EUROTRA NLP Basic research, VERBMOBIL	Basic research CICC, EDR, …

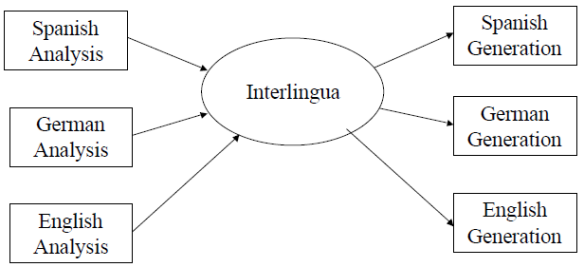
7. The MT-Triangle:



8. Transfer-System with 3 Languages:



9. Interlingua-System with 3 Languages:



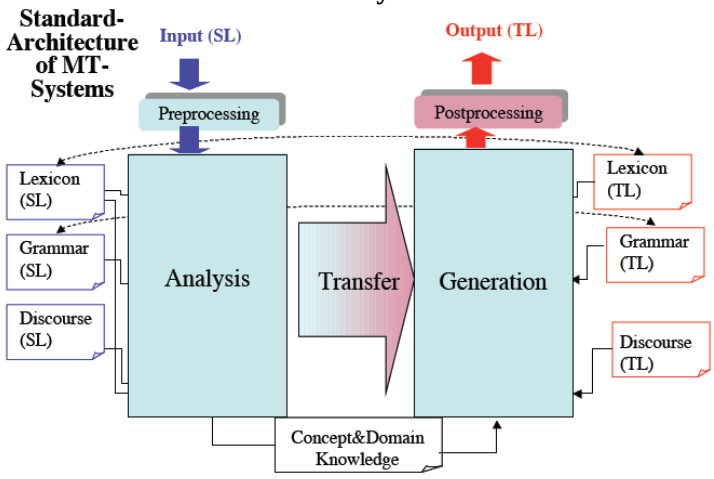
10. Interlingua-Systems:

- Each module is independent from all other analysis and generation modules
- Target languages have no influence on the analysis process.
- For a new language only 2 new modules have to be added
- "back-translation" possible (useful for system evaluation)
- Complicated representation even for languages belonging to the same family

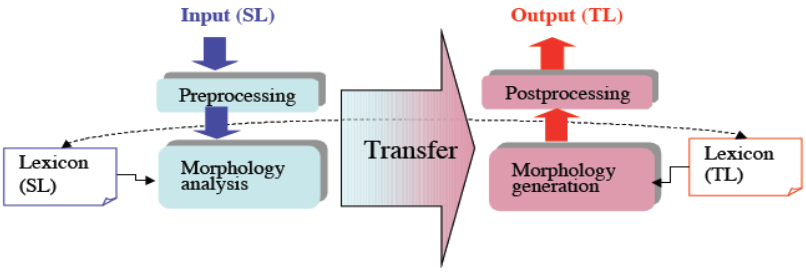
11. Transfer-Systems:

- Language-dependent
- For each new language a high number of new modules must be implemented (for **n** languages: **n(n-1)** modules)
- Straight-forward representation
- Local definition of similarities among languages.

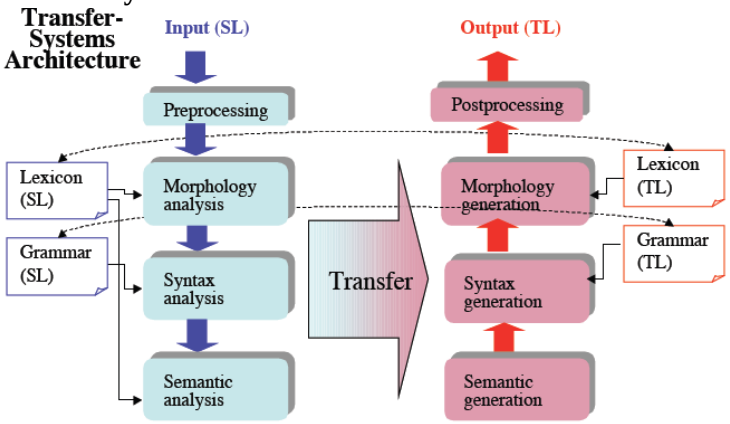
12. Standard-Architecture of MT Systems:



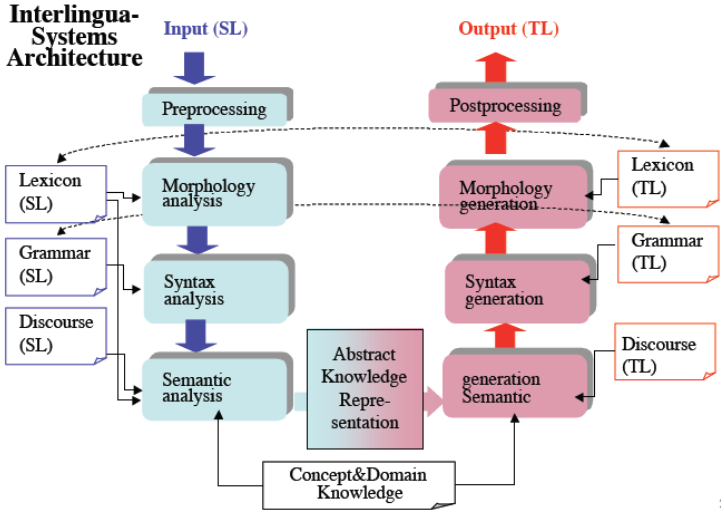
13. Direct System Architecture:



14. Transfer-Systems Architecture:



15. Interlingua-Systems Architecture:



16. MT-specific Pre-editing:

- Checking source texts for foreseeable problems for the system and trying to eradicate them
- It can include:

– Identification of names (proper nouns)
– Marking of grammatical categories of homographs
– Indication of embedded clauses
– Bracketing of coordinate structures
– Flagging or substitution of unknown words
– Extreme form: Reformulation of the text using a "controlled language" and a corresponding editor

17. Pre-editing—Controlled Language:

Adaptation of source texts to the vocabulary such constructions which the system can translate

The writers of texts for translation are restricted to:
– particular types of constructions
– the use of terminology,
– predefined meanings of every-day words

18. Post-Editing

Correction of the output from the MT-System to an agreed standard:
– Minimal for assimilation purposes
– Thoroughly for dissemination purposes

Interactive post-editing:
– The system alerts the editor of sentences or phrases which may be incorrectly translated (e.g. which contain an unresolved ambiguity, or a construction which could not be analysed
– It provides the option of correcting similar errors automatically throughout the text, once the editor has replaced a mistranslation

Linguistically intelligent word processors:
– Can spot some types of structural ambiguities
– Can generate alternative structures
– change automatically gender agreement in a whole phrase
– Insert automatically appropriate prepositions (e.g if discuss is changed to talk then about is inserted before the direct object)

19. Evaluation of MT-Systems:

In contrast to other software there is no "best solution" by human translators, which can be compared with the output of the system

i.e., for one input sentence there are many different correct translations

Quality measurement of an MT System depends on its purposes and on the requirements of potential users.

Possible participants in evaluation:
– Researchers
– Research sponsors
– Purchasers
– Translators

20. Evaluation strategies (*Black Box vs. Glass Box*)/ (*Test Suite vs. Test corpus*):

- **Black Box:**
  - MT system is seen as a black box, whose operation is treated purely in terms of its input-output behavior
  - Should not be conducted by the developers
  - Tests: functionality, volume of data handled, recovery situations
- **Glass Box:**
  - Components of the system are inspected as well a their effect in the system
  - Relevant to researchers and developers
  - Static analysis: checking the system without running it (automatic syntax and type checking by a compiler, manual inspection of the system, symbolic execution, data flow analysis)
  - Dynamic glass box requires running the program (e.g. trying the program on many logical paths and ensuring that every logical branch is executed at least once).
- **Test Suite:**
  - Carefully constructed set of examples, each testing a particular linguistic or translation problem (e.g. different lexical and structural differences)
  - Problem: it is assumed that the behaviour of a system can be projected from carefully constructed examples to real texts
  - Test suite evaluations are difficult to compare
- **Test corpus:**
  - An adequate corpus (for the domain of the system) is used as input
  - Problem: it does not test systematically all possible sources of incorrect translations, but considers the most frequent constructions
  - It is difficult to estimate the behaviour of the system for other types of text

21. Evaluation—Linguistic Quality measures:

- **Intelligibiltiy**—measures the *fluency* and *grammaticality* of the TL text, with concern for wether it faithfully conveys the meaning of the SL
- **Accuracy**—indicates how the translated text preserves the content of the source text. (a high intelligible sentence may not convey the meaning of the source text because of incorrect disambiguation)
- **Error analysis**: e.g. count the number of words inserted, modified, deleted and moved by a post-editor. However, deciding hat is an acceptable translation is subjective.

22. Evaluation—Software criteria:

- **Functionality**—determines the degree to which it fulfills the stated or implied needs of a user
- **Reliability**—if the system maintains its level of performance under specified conditions and for a specified period of time
- **Usability**—indicates the effort needed to use the software by a stated or implied set of users
- **Efficiency**—relationship between the level of performance of the software and the amount of resources used to achieve that level of performance under specified conditions
- **Maintainabiltiy**—effort needed to make specified modifications to the software
- **Portability**—indicates the ability of the software to be transferred from one environment to another.

23. Different Approaches to MT:

- Rule-based MT
- Knowledge-based MT
- Statistical-based -MT
- Example-based MT

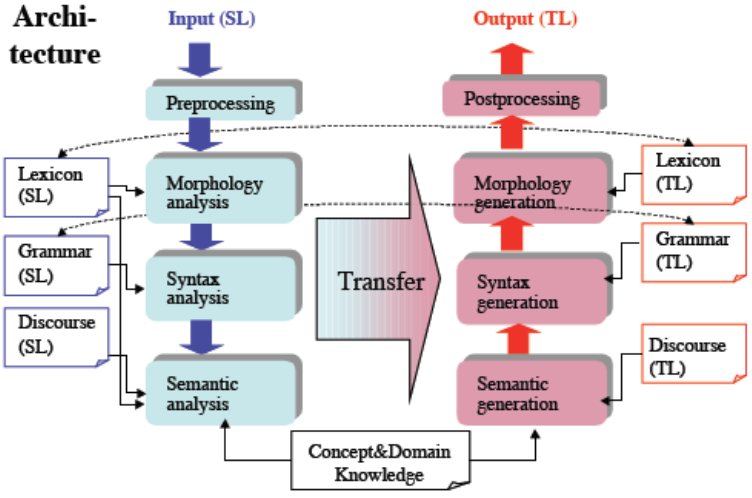
24. Other approaches to computer assisted translation:

- Machine Aided Translation
- Translation Memories

PART II

Rule based Machine Translation

1. Architecture:



2. Knowledge Sources:

- Bilingual (Multilingual) Lexicon
- Thesauri
- Grammars for SL and TL
- World /Domain Knowledge base
- Discourse memory

3. Thesauri:

- Are a particular form of lexicons and contain fixed expressions and their translations
- Expressions contained in such thesauri are replaced from the very beginning by their translations, and are no longer object of syntactic or semantic interpretations
  - ☑ *e.g.: {United States = Estados Unidos / Civil law = Código Civil}*
- ※ Sometimes abbreviations are also part of thesauri:
  - ☑ *e.g.: {Ud(s)=Usted(es) =You (politeness)}*
- ☆ Thesauri are domain specific

4. Limitations of morphological analysis:

- *No correct word-order:* the word-order can be solved by introducing transfer rules
- *Ambiguity:*
  - Lexical Ambiguity:
  - – Categorial ambiguity: the same word can belong to more than one
  - – Homography and Polysemy (the same word has more meanings) e.g. Bank (engl.)
  - – Translation ambiguity: e.g. the English leg can be translated in Spanish with pierna (human), pata (animal, table), pie(chair), etapa (of a journey)
  - Structural ambiguity or complicated syntactical problems

5. Lexical transfer:

- Consists usually of:
  - – Replacement of lexical elements in SL by their correspondents in TL
  - – If necessary correction of word-order {e.g. Adj Noun (English) => Noun Adj (Farsi)}
- ✪ Without problems when:
  - – There is a translation equivalent in TL
  - – Many to one translation i.e. more lexical items in SL are translated by the same item in TL

6. Problems of Lexical transfer:

- One-to many translation (one word in SL has different translations according to the context) {e.g. Wall (engl.) will be translated by *muro* (sp.) if it is outside and *pared* (sp.) inside.}
- Different translations depending on domain-specific/world knowledge {E.g. *biblioteca* (sp,) is translated into German by:

– Bibliothek if it belongs to an academic institution or is private
– Bücherei if it is a public library.

❄ Lexical gaps—single-word concepts in one language which can only be rendered by two or more words in the other {E.g. madrugó (sp.) = got up early (engl.)} This cannot be solved only by lexical transfer because in the English lexicon there is no entry "get up early"

7. Structural transfer:

⅄ is always necessary if the structure in SL can not be transferred to the TL, or it does not fit exactly due to semantic or stylistic rules

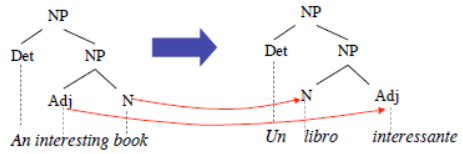
⌘ The deeper the analysis the more differences between languages disappear from the representation

▣ ***Solution:*** transfer rules

8. Syntactic transfer:

✺ Mapping between the surface structure of sentences: a collection of tree-to-tree transformations is applied recursively to the tree of the SL sentence in order to construct a TL tree

◈ The simplest form corresponds to word-order re-arrangements in lexical transfer:



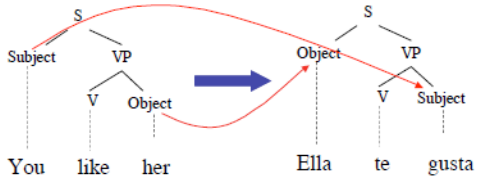
※ Tree-to-tree transformations:

– Recursive
– Top down process
– One side of the tree-to-tree transfer rule is matched against the input structure, resulting in the structure on the right-hand side

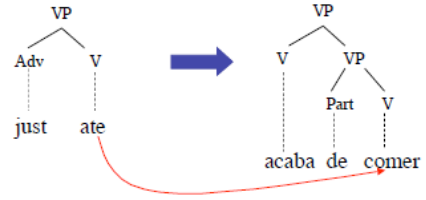
☆ Rules have to cover not only such simple cases but also:

– Thematic divergences
– Head switching
– Structural differences
– Lexical gaps
– Lexicalization
– Categorial divergences
– Collocational divergences

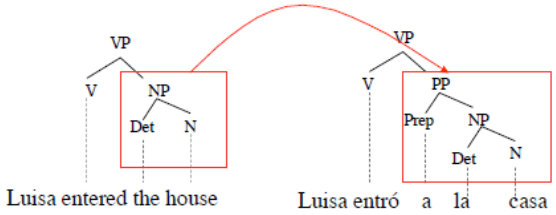
9. Syntactic Transfer: *Thematic divergences* *{Thematic divergences refer to changes in the grammatical role played by arguments of a predicate}*



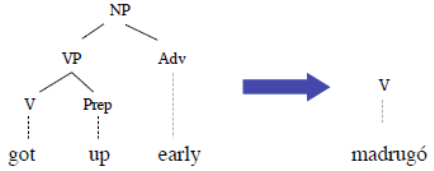
10. Syntactic Transfer: *Head Switching* *{The syntactic head of an expression in one language is translated as modifier, a complement or some other constituent in an other language}*



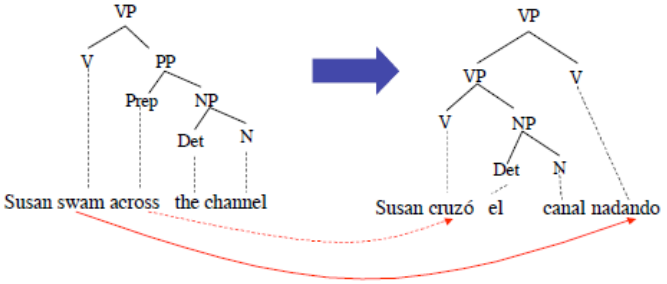
11. Syntactic Transfer: *Structural divergence* *{Different sub-constituents for the same constituent}*



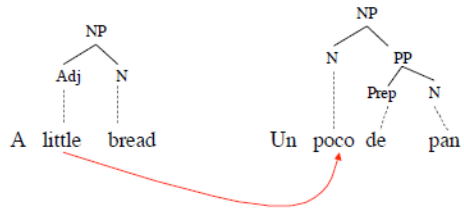
12. Syntactic Transfer: *Lexical gaps* *{For such cases special rules must be provided}*



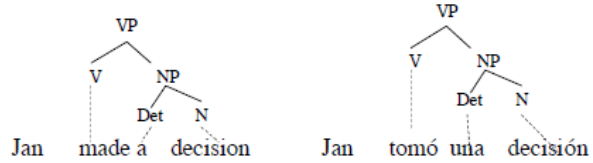
13. Syntactic Transfer: *Lexicalization* *{Languages distribute semantic content differently within a sentence}*



14. Syntactic transfer: *Categorial divergence* *{Although a one-to-one-translation exists, some words must be rendered via different syntactic categories. Sometimes this involves also head switching (I am hungry (engl.) => Tengo hambre (sp.))}*



15. Syntactic Transfer: *Collocational divergence* *{arise when the modifier, complement or head of a word is different from its default translation.}* [***Solution:*** list all combinations of relevant collocations and insert specific rules for each (take a walk = *dar una caminada*, be thirsty = *tener sed* etc.)]



16. Semantic Transfer: Semantic transfer interprets translation as a relation between language-dependent representations. The transformations are also recursive but they apply on the semantic representation.

17. Structural Transfer with Interlingua: *Tasks*:

- Content analysis from the Lexicon, Morphology, Syntax, Semantics, Pragmatics
- = Mapping of the Input on:
  - Presuppositions,
  - Objects (e.g. = Variables),
  - Relationships (e.g. Roles),
  - Quantifiers (e.g.= Negation, Number)
- Consistency check (e.g. Presupposition check)
- Semantic extraction
- Reordering of results in the generation phase

18. Generation in *Direct* Systems:

- Is reduced to:
  - *lexical substitution during the lexicon look-up*
  - *Local re-ordering*
- Generation is based on SL as much as possible
- Nothing else is changed, unless it is strictly necessary for the production of an acceptable target language expression.

19. Generation in *Transfer-based* Systems:

- Split into two modules:
  - *Syntactic generation*
  - *Morphological generation*

✍ **20.** In **syntactic generation** the intermediate representation which is the output of analysis and transfer is converted into an ordered surface-surface-structure tree, with appropriate labelling of the leaves with target language grammatical functions and features

✪ Main task of ***syntactic generation*** is to order constituents in the correct sequence of the target language

☑ e.g. For a sentence labelled as "passive" in the deep structure:

– *Syntactic generation creates a node for the auxiliary verb*
– *Labelled with the appropriate tense information*
– *Assign "past-participle" label to the main verb*

✍ **21. Morphological generation** processes the resulting surface structure:

– *Interprets strings of labelled lexical items for target output*
– *e.g: casa+pl = casas, ser+future+1stperson-sg. = seré*

✍ **22. Morphological generation** can usually be handled by a combination of general and special-case procedures, on a word-by-word basis.

✍ **23.** Generation in **Interlingua-Systems**:

❄ Additionally to the *syntactical* and *morphological* generation, there is a **semantic generation** component

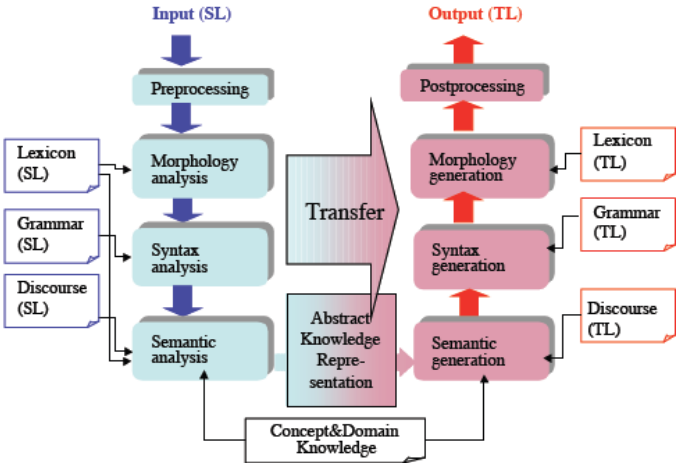
❁ The main task of the **semantic generation** is to find out which part of the interlingua expression should occur in the target sentence (e.g. not the existential presuppositions).

❁ The **semantic generation** produces as output a deep syntactic structure (i.e. a structure which has syntactic and semantic information, but is TL dependent)

PART III

Knowledge Based Machine Translation

✍ **1. Standard Architecture**:



2. In a sense, systems that use terminological material (in a systematic order according to the domain), can be called knowledge based systems. However, the ontological knowledge (conceptual ordering) of the field is not declarative, but implicit in the ordering of the terminology (nomenclature). The ontology is not visible directly.

3. Knowledge Layers: Knowledge used in an MT system may be:

- Conceptual knowledge ("ontology", "upper model")
- World knowledge (chemical laws, e.g.)
- Factual knowledge (situational knowledge) about the actual state of affairs

4. Three Examples of the use of **knowledge** in MT:

- KBMT 1989
- DBR-MAT 1999
- Verbmobil 2001

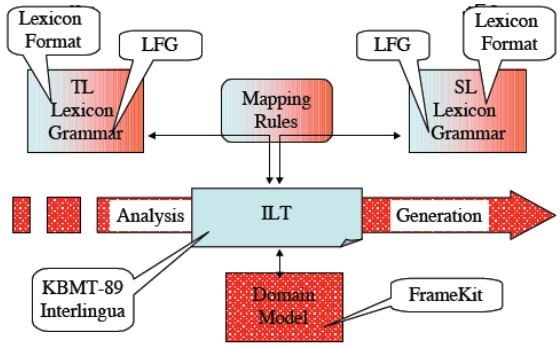
5. Assumptions behind **KBMT:**

- One "functionally complete" meaning representation can serve for translations to a number of languages,
- No total representation of human understanding of a text is necessary for machine translation,
- Applicable to relatively unambiguous, e.g. technical documents.

6. Basic Components of a **KBMT** System:

- An ontology of concepts ("domain model", "ontology")
- Source language (SL) lexicon and grammar for the analysis process
- Target language (TL) lexicon and grammar for the generation processes
- Mapping rules between the Interlingua and SL/TL syntax.

7. Knowledge Bases and Languages:



8. KB: *Frames with linguistic and nonlinguistic knowledge*:

- The KBMT-89 ontology contains
  - Objects
  - Events
  - Properties of objects or events
  - Relations
  - Attributes
- Concepts are linked to others by relations. Each concept has attributes which specify value sets. Value sets contain only literals (i.e. no concepts).

9. Critical **problems** of **knowledge-based** systems are still

1. The huge effort to build up knowledge bases,
2. A practical definition of the size (" coverage") of the knowledge base, and
3. The choice of the representation language and its necessary logical/formal properties.

10. Advantage of Knowledge Bases

- Using knowledge bases the developer definitely knows, what is represented where, although he cannot predict, what can be derived with the inference rules. With implicit and procedural (local) representations there is no method to check multiple representation.
- Declarative knowledge sources, are global, can be maintained in isolation, can be exchanged and may be used in other inference machines or grammars. It even can be used in other systems than translation systems.

11. The Semantic Web Idea:

- This basic idea of declarative and modularized knowledge has become very important since the famous paper of *Berners-Lee* in 2001 on the "semantic web".
- The **Semantic Web** is the abstract representation of data on the World Wide Web, based on the RDF standards and other standards to be

defined. It is being developed by the W3C, in collaboration with a large number of researchers and industrial partners.

- The **semantic web** will serve as one (or several) ontology(ies) to which all WWW objects refer and which can be used consequently for web operations like data mining, information extraction, summarization, etc" [and translation!]
- The first obvious result of the **semantic web** activities for translation is, that widely accepted ontologies of specific domains can be used as knowledge bases for machine translation. This solves point (1) to (3) of the above mentioned list of problems.

PART IV

Corpus-based Machine translation

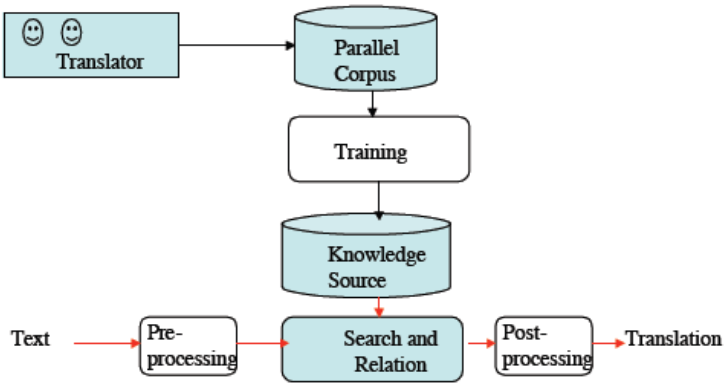
1. General Principles:

- The linguistic phenomena in both languages as well as the transfer rules are no longer linguistically described but derived automatically from a parallel corpus.
- First an aligned corpus is built
- Next step is a training phase, in which are calculated the connections between elements in the source language as well as in the target language (sometimes the results are calle „knowledge sources")).
- The translation is the result of 2 processes:
  - *A search process (of elements in the source language)*
  - *A best-evaluated relation with a target expression*

2. There are 2 types of **corpus-based** MT systems

- **Example based MT**: The translation of a source text is based of translation examples in the database
- **Statistical MT:** the alignment information from the corpus is used for the training of a statistical translation model

3. Generical Architecture of a **corpus-based** MT:



4. Aligned Corpus:

- A *parallel Corpus*:
  - – Is a collection of texts in at least 2 languages. It is extremely important that the content is the same for both texts.
  - – Examples: Official Documents from EU, Newspapers in Countries with more than 1 official language
  - – Contains markers (*zags*) for content-identical elements (Sentences, Paragraphs) in Texts
- ✪ The parallel aligned Corpus has to be adequate for the translation domain.
- ❄ When searching such corpora the main problem is, that 1 chunk in the source has more than 1 translation in the target language and the choice is made according to the context.

5. Alignment-Methods:

- Manual:
  - – Extreme time consuming, because for real applications the corpus has to be really big.
  - – Specialists with very good knowledge in both languages are needed
- ☸ Automatic with help of statistical procedures:
  - – e.g. length-based methods (number of words in the source and target text has to be close one to another)
  - – Difficult to identify at the word-level, because for e.g. in: [*Haben Scharfe Kritik geuebt* => have strongly criticized] The POS chaneg and the semantic combinations are different.

6. Statistical MT: *Principles*

- Given:
  - – A source sentence (e.g. in German.): D = d_1, ...d_i,...., d_n (d_i are the words) which has to be translated into a sentence (in English for e.g..) E = e_1, ... e_i, ..., e_m.
  - – A parallel aligned German-English corpus
- ⌘ Between all translation possibilities it is searched the one with the highest probability. This means mathematically:

$$\hat{e}^m = \arg\max_{e^m} \left\{ \Pr\left(e^m \middle| d^n\right) \right\}$$

7. Statistical MT: *Model*

- ✠ ***Das Source-Channel Model*** (used very often). Following decomposition is used: *(Both models are dependent of parameters, which are calculated in the training phase)*

$$\Pr\left(e^m \middle| d^n\right) = \Pr\left(e^m\right)\Pr\left(d^n \middle| e^m\right)$$

Language model - give the probability that e^m is a correct english sentence

Translation model

Give the probability that in the corpus a sentence e^m will be find which is the translation of d^n.

- ✺ ***Direct Maximum Entropy Translation Model*** (The original probability is calculated directly, following different translation features (mathematically is a function with parameters):

$$\Pr\left(e^m \middle| d^n\right)$$

- 🔲 ***Alignment Model***: [A new parameter is introduced, which models the alignment mapping. Here features like *Fertility* and *Distortion* are considered]

8. Advantages of Statistical MT:

- ◈ Use no linguistic knowledge (as long as the alignment of the corpus is done automatically)
- ☆ Loose dependencies between constituents can be modelled better with statistical models as with rules
- ⌖ It is especially indicated to be used in embedded systems e.g. in Speech Systems, where a language model already is defined (for the speech recognizer)

9. Well-known **problems with Statistical MT:**

- ⌗ New field, there are few systems which can be evaluated. (Verbmobil, Tanslation of Canadian parliament debates)
- ※ Exceptions can be trained difficult
- 🖹 Morphology:
  - – Inflected forms of the same word are treated as not-related words. e.g. the Word *diriger* in French is translated with *führen* or *leiten* in German. For each one of the 39 inflected forms of the word the model has to be trained.

✪ Not-local dependencies are difficult to be trained. The System produces usually correct word-translations but in an incorrect order

❄ Probabilities for rare words are not to be trusted.

❁ The models are very sensible to data-changes.

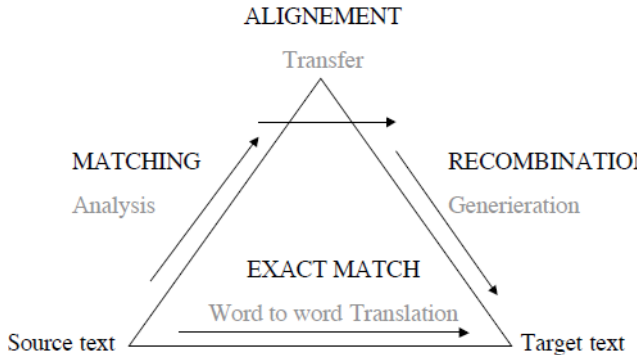
10. General Principles of **EBMT**:

❁ A parallel Corpus is used

⅄ Part of the input text are compared with source chunks in the corpus

⌘ The translation of the founded parts are put together and form the translation.

11. Translation pyramid for **EBMT**:



12. Length and Size of Examples:

✠ The size of the example database varies between some hundreds and 800000 sentences.

✺ The bigger is the Database the better works the system

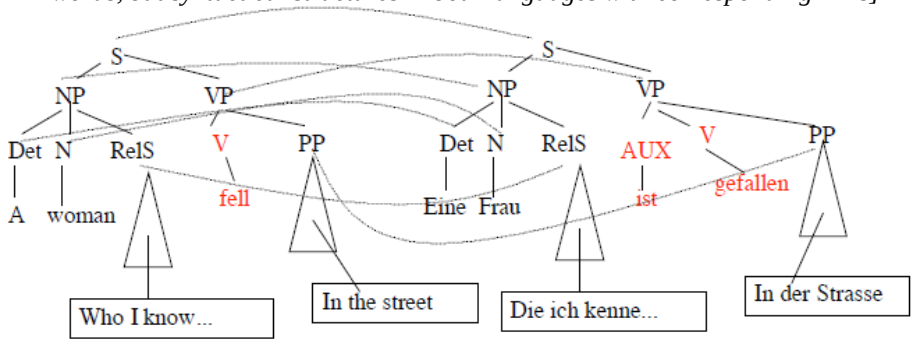
▣ There is no ideal length for the examples:

– *The longer the examples, the lower the chance for a match*

– *The shorter the example the bigger is the chance to have some ambiguities*

◈ Usually the standard unit for the examples is a sentence

13. EBMT with linguistic knowledge: *[The Translation patterns are not words, but syntactical structures in both languages with corresponding links]*



14. What are **Translation Memories**?

- ☆ Translation Memories TM are big databases with translation examples
- ◎ They contain also a search mechanisms
- ⌖ Example: TRADOS—a TM-System for all old 12 EU languages

15. EBMT and **Translation Memories:**

- ⌗ The search mechanism (analysis) is the same.
- ※ TM have no combination phase.
- 🗎 Translation Memories are helpful for translators but do not replace them, therefore the output is a list with translation alternatives for parts or complete sentences.

16. EBMT and **Statistical MT**:

- ⁂ There are trials to combine the two approaches
- ✹ The idea is exceptions which cannot be statistical modeled or very often expressions to be filtered through example search.

17. Comparison of **linguistic** and **empirical methods**:

- ✪ In Verbmobil-System (German-English-Japanese Speech-to-Speech System) were 5 MT approaches implemented, 1 transfer-bases, 1 statistical and 1 example based.
- ❄ After the evaluation of the mistakes:
  - Semantic Transfer 62 %
  - Example based MT 35%
  - Statistical MT 29%
- ❁ Most problems of the empirical approaches are due to:
  - ***Word order***: the target language model is not trained accordingly
  - ***Disambiguation***: Very difficult for prepositions which are translated according to the context
  - ***No partial translation***: statistical translation process only sentences. If the speech recognizer contains mistakes or "false starts" the quality of the translation decrease rapidly.
  - ***Problems with verb particles***: in this case is a morphological preprocessing needed, but this increases very much the processing time

18. Corpus-Linguistics and Statistics for Linguists Links:

- http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm - Course on Corpus Linguistics
- www.coli.uni-sb.de/~christer/stat_cl.ps Linguists guide to statistics

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 (۱)

شامل مهمترین نکاتِ چهار 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