

New Trends in Machine Translation

Celia Rico Pérez et Aurora Martín de Santa Olalla Sánchez

Volume 42, numéro 4, décembre 1997

URI : <https://id.erudit.org/iderudit/003822ar>

DOI : <https://doi.org/10.7202/003822ar>

[Aller au sommaire du numéro](#)

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)

1492-1421 (numérique)

[Découvrir la revue](#)

Citer cet article

Rico Pérez, C. & de Santa Olalla Sánchez, A. M. (1997). New Trends in Machine Translation. *Meta*, 42(4), 605-615. <https://doi.org/10.7202/003822ar>

Résumé de l'article

Le but du présent article est d'offrir un aperçu des méthodes actuelles de TA basée sur le corpus. Il présente d'abord la linguistique basée sur le corpus, discipline qui a donné naissance à ces nouvelles méthodes. Il traite ensuite d'aspects tels que l'annotation et l'alignement des corpus. L'article fait également le point, en termes généraux, sur la TA basée sur des exemples et la TA basée sur les statistiques. Finalement, il montre l'adéquation de ces modèles à une approche réaliste de la traduction.

NEW TRENDS IN MACHINE TRANSLATION

CELIA RICO PÉREZ AND AURORA MARTÍN DE SANTA OLALLA SÁNCHEZ
Universidad Europea de Madrid and Universidad Alfonso X El Sabio, Madrid, Spain

Résumé

Le but du présent article est d'offrir un aperçu des méthodes actuelles de TA basée sur le corpus. Il présente d'abord la linguistique basée sur le corpus, discipline qui a donné naissance à ces nouvelles méthodes. Il traite ensuite d'aspects tels que l'annotation et l'alignement des corpus. L'article fait également le point, en termes généraux, sur la TA basée sur des exemples et la TA basée sur les statistiques. Finalement, il montre l'adéquation de ces modèles à une approche réaliste de la traduction.

Abstract

This article offers a general overview of current methods in corpus-based MT. It first presents corpus linguistics as the discipline from which new research methods have sprung up and deals with aspects such as corpus annotation and alignment. It introduces both example-based and statistics-based MT from a general perspective, and finally advocates these methods as a realistic approach to MT.

INTRODUCTION

Over the past few years research in Machine Translation (MT) has focused on the techniques and studies developed for corpus processing with a view to using these recent experiences as the starting point for MT purposes. This is not a new idea for, as early as 1949, Warren Weaver suggested using statistical techniques, and hence data extracted from real texts, for the task of translation. The direct consequence of this corpus approach is a growing number of interesting works which have opened new fields of research by drawing on the advances and possibilities of compiled corpora, and belong to what is now known as corpus MT.

Previous MT research, usually referred to as 'traditional' or 'linguistic', has concentrated on the development of rule-based systems where the standard data structure is the *tree* and the standard operations are various kinds of *tree-to-tree transductions* or, more recently, systems based on *feature structures* and *unification* (Sadler and Arnold 1992). Traditional systems usually encode lexical, syntactic, semantic and, rarely, pragmatic knowledge as a set of 'linguistic' rules for analysis, transfer and generation. These systems have shown interesting results and some have become successful commercial tools. Nevertheless, MT research is far from settled once and for all, and new techniques have to be developed.

In an attempt to explore new fields of research, corpus MT adopts an empirical point of view and turns to corpora as the source of information for extracting real data. The motivations behind this interest are, on the one hand, to reach consistency in describing knowledge and thus avoid false intuitions when building rules, and on the other, to use previously translated texts as the source of knowledge.

There are two main tendencies in corpus MT: example-based MT and statistics-based MT. Both approaches use corpora but with different methodologies for extracting knowledge from real texts. Basically, example-based MT encodes knowledge from corpora

as translation patterns and then looks for the best translation match for the source language text, whereas statistical MT assigns a probability of translation to every pair of sentences.

The aim of this article is to offer a general overview of current methods in corpus-based MT. It first presents corpus linguistics as the discipline from which new research methods have sprung up and deals with aspects such as corpus annotation and alignment. Next, it introduces both example-based and statistics-based MT from a general perspective, and, finally, advocates these methods as a realistic approach to MT.

The article is, therefore, aimed at those who seek an introduction to new MT methods and latest advances in the field. It directs the reader to specific bibliography in corpus processing, example-based techniques and statistics-based methods and helps him to understand the fundamental principles behind these techniques.

REFERENCE CORPORA AND NATURAL LANGUAGE PROCESSING

In the last three decades corpus computational linguistics, the linguistic branch related to the study of language from large textual corpora, has developed into a discipline in its own right, offering support to other linguistic branches too.

The origins of corpus-based linguistics precede the use of large textual corpora and the use of computers to store data. They must be traced back to the era of American post-Bloomfieldian structural linguistics, when texts (written and spoken) were considered as the primary and only source of information for linguistic research.

There is a virtual discontinuity, however, between the corpus linguistics of that era and the later variety of corpus computational linguistics. The main difference lies in the use of computers to store this large amount of data, with all the advantages that this involves in terms of size/capacity and possibilities of information access and retrieval. As G. Leech (1991) points out, this innovation involves the following three aspects:

1. The value of the corpus as a source of systematic retrieval of data.
2. The value of the corpus as testbed for linguistic analysis or hypothesis.
3. The value of the corpus as a methodology for building robust natural language systems.

A computational corpus is now defined as a collection of machine readable texts, a textual archive or database integrated in an information storage and retrieval system.

Parallel Corpora: Annotation and Alignment

The corpus designation can be applied to collections of texts with very different characteristics: written or spoken, sample texts or complete texts, sublanguage texts or general texts, in a number of languages. According to these characteristics, a corpus typology can be set up. Thus, corpora can be mono- or multilingual, and among the last group, we distinguish parallel corpora as the ones used for MT.

A parallel corpus is a collection of texts, each translated into one or more languages. The simplest case involves only two languages: one corpus is an exact translation of the other. However, we can find some parallel corpora which contain translation into several languages. It is also important to mention that the direction of the translation need not be constant, so that some texts in a parallel corpus may have been translated from language A to language B and others the other way around. The direction of the translation may not even be known.

Examples of parallel corpora are found in the area of communication in multilingual societies, such as the United Nations, NATO, the EU and officially bilingual countries such as Canada.

Corpora Annotation

Computational corpora must have some kind of annotation if we are going to make linguistic use of them, that is to say, if we are going to use them in the construction of grammars, dictionaries or in corpus based MT.

Annotation is, broadly speaking, a process of making explicit what is conjectural or implicit, a process of directing the user as to how the content of the text should be interpreted.

Linguistic annotation is the practice of adding interpretative linguistic information to an existing corpus of spoken or written language, by some kind of coding attached to, or interspersed with, the electronic representation of the language material itself.

Linguistic information has many different degrees of delicacy or granularity. We distinguish here a two-level distinction: encoded or tagged corpora and analysed corpora. Encoded corpora mainly refers to part-of-speech tagging, which allows simple syntactic searches to be carried out. At the end of this stage corpora are characterised and ready for the analysis task. On the other hand, analysed corpora normally include, in addition to some form of tags, information about 'higher level' analysis such as brackets identifying phrases of various types (nominal groups, prepositional groups, etc.), labelled parse-trees for each sentence, etc.

Annotation Standards

Researchers, language engineers and linguistic technology inventors have recently become aware of the ideas of reusability and interchangeability in the creation and development of linguistic technology products satisfying different user needs. Annotation standards have been set as proposals (most of them within the European framework) for text encoding and analysis in different languages.

Encoding and analysis proposals use what we call a mark-up language. By mark-up language, we mean a set of mark-up conventions used together for encoding texts. A mark-up language must specify what mark-up is allowed, what mark-up is required, how mark-up is to be distinguished from text, and what the mark-up means. The Standard Generalised Mark-up Language (SGML) provides the means for doing the first three; proposals such as the one advanced by the Text Encoding Initiative (TEI) Guidelines provide the last one.

SGML is an international standard (ISO 8879) for the description of marked-up electronic text. More exactly, SGML is a metalanguage, that is, a means of formally describing a language, in this case, a mark-up language.

In 1987, with funding from the U.S. National Endowment for the Humanities, the Association for Computers and the Humanities (ACH) organised a workshop to investigate the possibility of developing an encoding standard for machine-readable texts. As a result of this, TEI was constituted as a four year international research project whose goal is to provide guidelines for the encoding of electronic texts.

Since this date, TEI has created different workgroups or committees directed by a *Steering Committee*. Each workgroup was set up initially with a specific task. From the moment the groups were established, publication of drafts, meetings and reviews have taken place. At the moment there are public electronic lists where information about ongoing work can be obtained. The last TEI publication (TEI P3) was issued in 1994.

There are two European projects related to TEI and encoding standards: The Network of European Reference Corpora (NERC) and The Expert Advisory Group on Language Engineering Standards (EAGLES).

NERC is a joint project of six European countries aimed at the constitution of a European network for corpora in order to serve the various European Language engineering

needs, through a European coordination of national efforts. One of its tasks is to define textual representation for European corpora, both written and spoken.

EAGLES is an initiative of the European Commission launched in February 1993, within the EC Directorate General 13th Linguistic Research and Engineering programme, as a LRE project. The aim of EAGLES is to accelerate the provision of standards for:

- very large-scale language resources (such as text corpora, computational lexicons and speech corpora);
- manipulating such knowledge, via computational linguistic formalisms, mark up languages and various software tools;
- assessing and evaluating resources, tools and products.

Coordination is being carried out by the *Consorzio Pisa Ricerche* (Italy), while five working groups execute the detailed work. The one in charge of Text Corpora is placed at the *Instituto Cervantes* in Spain.

Up to now the three encoding proposals have concentrated mainly on morphosyntactic aspects, that is to say, on the encoding of explicit formal mark-up that implies a specific grammatical behaviour. A morphosyntactic mark-up will tell us that a word as the Spanish *Juan* is a masculine, proper noun (N4MS), the Spanish verb *cantan* is a third plural person, simple present of an auxiliary (VDR3P5), or that ? is a closing punctuation mark (PUC). It is at this level that we can most easily reach a consensual proposal containing all the common or specific features of the European languages.

While the TEI Linguistic Committee has proposed standards both for the formalisms and the contents of the tags, NERC has devoted itself to the definitions of contents. Finally, EAGLES has not only given standards both for the formalisms and the contents of the tags but has also proposed different levels of standardisation, certain optional lexical or lexico-semantic features divided into two groups: those which are specific to certain tasks or applications and those which are specific to certain languages. EAGLES proposes an Intermediate Tagset that can be used as a language-neutral representation of a set of attribute-value pairs, based on their word categorisation. For example, the order of attributes for the noun category (C=N) will be the following: Proper (P=4) or Common (P=5); Masculine (G=M), Feminine (G=F) or Invariable (G=6); and Singular (N=S), Plural (N=P) or Invariable (N=6).

They all agree on the representation of morphosyntactic analysis in a feature-structure formalism. Tags or labels have an atomic form represented by a string of characters. They are a composite which permits recovery of their attribute-value structure because of the position of each individual value. Their content includes not only the categories or grammatical class the words belong to but also the common and specific features of each individual form.

The state of the art in the encoding of corpora in European language is such that morphosyntactic tagsets have been defined for almost every language. Among them, English is the one with the greatest number of encoding proposals not only for the morphosyntactic level but also for the lexical or lexical-semantic one.

Concerning the encoding or tagging of parallel corpora, we would point out that EAGLES and NERC proposals have been applied, tested and evaluated in European projects like LRE CRATER and MULTEXT.

Alignment

Alignment is the technique of identifying correspondences between sentences in one language and sentences in the other language. The input is a pair of texts such as Table 1. The output identifies the alignment between sentences (Table 2).

Although alignment techniques began in 1988 with Brown, Lai and Mercer at IBM and Catizone, Russell and Warwick at ISCO, at the time of writing we have information about two more algorithms: those of Gale and Church at AT&T and Martin Kay and Martin Röscheissen at Xerox Palo Alto Research Center.

Both Brown *et al.* (1991) and Gale and Church (1993) use methods based on the observation that the length of the text unit is highly correlated to the length of the translation of this unit. Brown *et al.* measure words per unit, Gale and Church measure units in number of characters. On the other hand, Catizone *et al.* and Kay and Röscheissen (1993) present a lexical approach starting from correspondences between words based on the similarity of their distributions.

Gale and Church (1993: 78) describe a method and a program for aligning sentences based on a simple statistical model of character lengths. The program uses the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. A probabilistic score is assigned to each proposed correspondence of sentences, based on the scaled difference of lengths of the two sentences and the variance of this difference. The probabilistic score is used in a dynamic programming framework to find the maximum likelihood alignment of sentences.

Table 1

Input to alignment program.

English	Spanish
One of the most significant findings to come out of all Microsoft's research is that kids prefer a way of manipulating the objects on the screen that's different from what we're used to. Traditional Windows and Macintosh programs are arranged in noun/verb format: You select an area of text or an object, then you choose an action to perform on it. Microsoft's hundreds of test subjects preferred the opposite: You choose a tool, then drag it across the object.	Uno de los hallazgos más importantes de la investigación de Microsoft es que los niños prefieren una forma de manipular los objetos diferente a la que nosotros estamos acostumbrados. Los programas clásicos de Windows y Macintosh se organizan en un formato nombre/verbo. Se selecciona un área del texto y a continuación la acción que se llevará a cabo. Los cientos de sujetos que pasaron las pruebas de Microsoft eligieron la opción contraria: se elige la herramienta y se arrastra a lo largo del objeto.

Table 2

Output from alignment program.

English	Spanish
One of the most significant findings to come out of all Microsoft's research is that kids prefer a way of manipulating the objects on the screen that's different from what we're used to. Traditional Windows and Macintosh programs are arranged in noun/verb format: You select an area of text or an object, then you choose an action to perform on it. You select an area of text or an object, then you choose an action to perform on it. Microsoft's hundreds of test subjects preferred the opposite: You choose a tool, then drag it across the object.	Uno de los hallazgos más importantes de la investigación de Microsoft es que los niños prefieren una forma de manipular los objetos diferente de la que nosotros estamos acostumbrados. Los programas clásicos de Windows y Macintosh se organizan en un formato nombre/verbo. Se selecciona un área del texto y a continuación la acción que se llevará a cabo. Los cientos de sujetos que pasaron las pruebas de Microsoft eligieron la opción contraria: se elige la herramienta y se arrastra a lo largo del objeto.

The method is also fairly language-independent. Both English-French and English-German data were processed using the same parameters. If necessary, it is possible to fit the six parameters in the model with language-specific values, though, so far, it has not been necessary to do so.

It has been used in a trilingual corpus of economic reports issued by the Union Bank of Switzerland (UBS) in English, French and German. The method correctly aligned all but 4% of the sentences. There were more errors on the English-French subcorpus than on the English-German subcorpus.

To further research on bilingual corpora, a much larger sample of Canadian Hansards (approximately 90 millions words, half in English and half in French) has been aligned with the align program as a first step for MT. The results have been better in this case.

Kay and Röscheisen present an algorithm for aligning texts with their translations that is based only on internal evidence. The method depends on no information about the languages involved beyond what can be derived from the texts themselves. The plan rests on a relationship between words and sentence alignments arising from the observation that a pair of sentences containing an aligned pair of words must themselves be aligned.

In this case it is relatively easy to establish a correspondence between words such as proper nouns and technical terms. The main difficulty of this algorithm is to decide just which words in an original are responsible for a given one in a translation, and in any case some words apparently translate morphological or syntactic phenomena rather than other words, i.e. prepositions and conjunctions.

Up to now and as far as we know, this technique has been applied to an article from *Scientific American* and its German translation in *Spektrum der Wissenschaft*. We have no evaluation of results about it.

PARALLEL CORPUS AND MT

Once parallel corpora have been encoded and aligned, they are ready to be used for MT purposes. The information about language correspondences is explicitly shown in the aligned parallel texts and the next step is to use this information as the source for translation of new texts in the same languages.

As we have already mentioned, there are two ways of exploiting this knowledge. One method uses translation examples as patterns; the other uses frequencies of translation extracted from the aligned texts to calculate translation probabilities of new texts.

Example-based MT

We can trace the first experiences in example-based MT back to Nagao's (1984) work in this area. With the basic idea of MT as an imitation of previously translated text, example-based MT aims at producing translations by searching first a parallel corpus for the best possible match between the source language text to be translated (SLT), and the source language text stored beforehand, i.e. the source language model text (SLMT). Once the match is found between the SLT and the SLMT, the translation for the SLT will be that corresponding to the SLMT in the bilingual corpus, i.e. the target language model text (TLMT). See Figure 1 for a model of this process.

The whole process is, obviously, not as straightforward as it may seem. It involves several tasks which have to be carefully addressed: building a translation model, determining the translation units (words, phrases, sentences), deciding on translation equivalents and establishing the degree of similarity between the model and the source input text — a process usually known as *distance calculation*.

Building Translation Models

Before the system can actually perform any translation it needs a model establishing the patterns that have to be followed. This model is made up of contextual equivalencies between the source language and the target language. These equivalencies are extracted directly from the parallel corpus and encode explicitly or implicitly all linguistic information needed for translation.

Building the model involves two fundamental tasks: a) extracting knowledge; b) formalising this knowledge as patterns. The first task, that of extracting knowledge, represents one of the fundamental problems not only for MT but also in any other area of Natural Language Processing, it deals with issues which are close to Artificial Intelligence.

Basically, we can establish two ways of extracting knowledge: manually and automatically (or rather, semi-automatically). Manual extraction requires the activity of a group of experts who analyse the parallel corpus and decide on rules and create templates. This technique is the most frequent, although it has some obvious drawbacks: it involves the work of extensively trained human experts who master both languages, who must know the system components to update rules, and who may write slightly different rules according to different criteria. As a result, maintenance of this type of rules is difficult (Almuallim *et al.* 1994).

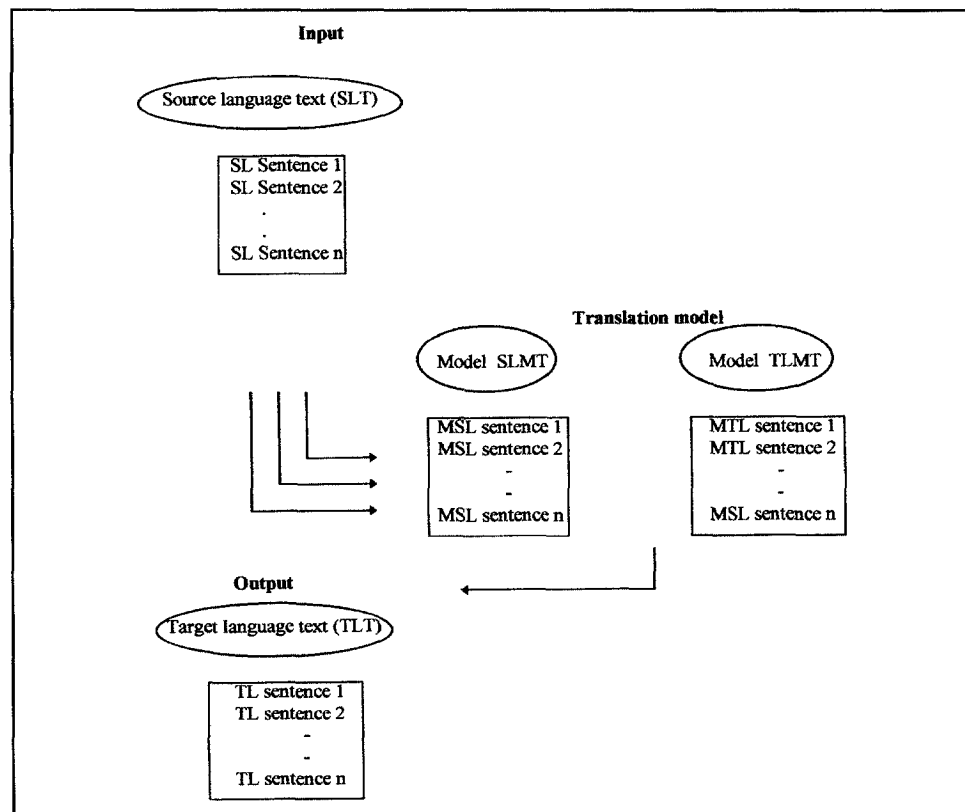


Figure 1:
Translation process in example-based MT

In order to overcome these difficulties, some attempts have been made at automatically extracting translation templates. These experiments implement rather semi-automatic techniques, as full automation is at the moment too challenging. Algorithms used include AI techniques (Almuallim *et al.* 1994), or linguistic analysis through the use of dictionaries and parsing techniques both for identification of bilingual semantic features correspondences and for coupling structures (Kaji *et al.* 1992; Sadler 1990). Whether manual or semi-automatically extracted, these rules or translation templates, encode source variables and conditions to be matched with target variables.

It is important to stress the contextual character of translation equivalents since it is this quality that allows source language knowledge to be passed directly onto target language knowledge. In fact, as many studies have already shown, encoding world knowledge is one of the stumbling blocks not only for MT but for any field of NLP research. It is the case that morphological, syntactic and semantic information can be formalised in rules which the MT system uses straightforwardly for analysis, transfer and generation, but when we turn to encoding world knowledge, the task is not easily addressed.

We believe example-based techniques can help to overcome this problem because the translation model extracted from the bilingual corpus establishes direct equivalencies between source and target language in context, thus avoiding false equivalencies taken out of their context.

Calculating the Distance

Once the translation templates are encoded, the next step in the translation process is to find the best possible match between the text to be translated and the template. This process is usually known as *distance calculation* and it involves estimating the degree of similarity between the source pattern and the input text.

There are several ways in which similarity can be established but it usually involves finding the best candidate according to closeness of semantic attributes in a thesaurus (Sumita and Iida 1991) where thesaurus codes correspond to semantic attributes. It can also be established as word-based metrics, comparing individual words in terms of their morphological paradigms, synonyms, hyperonyms, hyponyms, part of speech tag (Nirenburg *et al.* 1993); or as syntax-rule driven metrics, trying to capture similarity of two sentences at the syntax level.

Once distance has been calculated, the system selects the most appropriate rule or template according to this result, usually based on the definition of a score of translation which reflects the correctness of the translation (Sato and Nagao 1990).

Statistics-based MT

The idea of applying statistical methods to MT was brought forward by Brown around 1988. The aim of statistical MT is, generally speaking, to calculate the probability of an example target text as the translation of a source text and offer it as the basis for further translation. In Brown's words, the main idea of this technique is not to imitate but to translate through estimating the probability that, given any two sentences in a source and a target languages, one is the translation of the other. Figure 2 shows a model of this process.

The motivation behind this method is the same as in example-based MT: using previous experience as the source of information. While example-based MT's models consist of conditions and variables for source and target patterns, statistics-based MT uses probabilistic models of both languages.

The parameters of these models are estimated automatically from a large database of source-target sentence pairs and probabilities of translation are derived for converting source language words and expressions into target words and expressions.

Statistics-based methods have been mainly applied by IBM in a project which uses the Canadian Hansard corpus of parliamentary debates as the basis for translation from French into English. This project employs statistics-based approaches developed in speech recognition. The central strategy consists, first, of estimating the parameters for the translation model, a task which is determined automatically by the alignment of sentences in the two languages, and, secondly, of calculating the probability that a word corresponds to zero, one or more words in the target language.

In order to establish the parameters, two probabilistic models are needed: a model of the language and a model of translation (Brown *et al.* 1990). The model of the language establishes the probability of occurrence of a single word given all the words that precede it in a sentence, i.e. its history. This model takes into account the relative position of the word and depends on word positions and sentence distribution in the corpus.

As for the model of translation, it involves estimating the probabilities of a word in the source language producing a particular word in the target language. Not all words can be paired in both languages, as sometimes there are no words for the source language which correspond to a target language word, and vice versa. It is necessary then, to introduce a *fertility* measure for each word in the source language producing target language equivalences. This fertility measure establishes the probability for each source language word producing zero to some moderate limit of words in the target language. The actual translation is generated by selecting that sentence *S* in the source language for which the probability of translation is maximum.

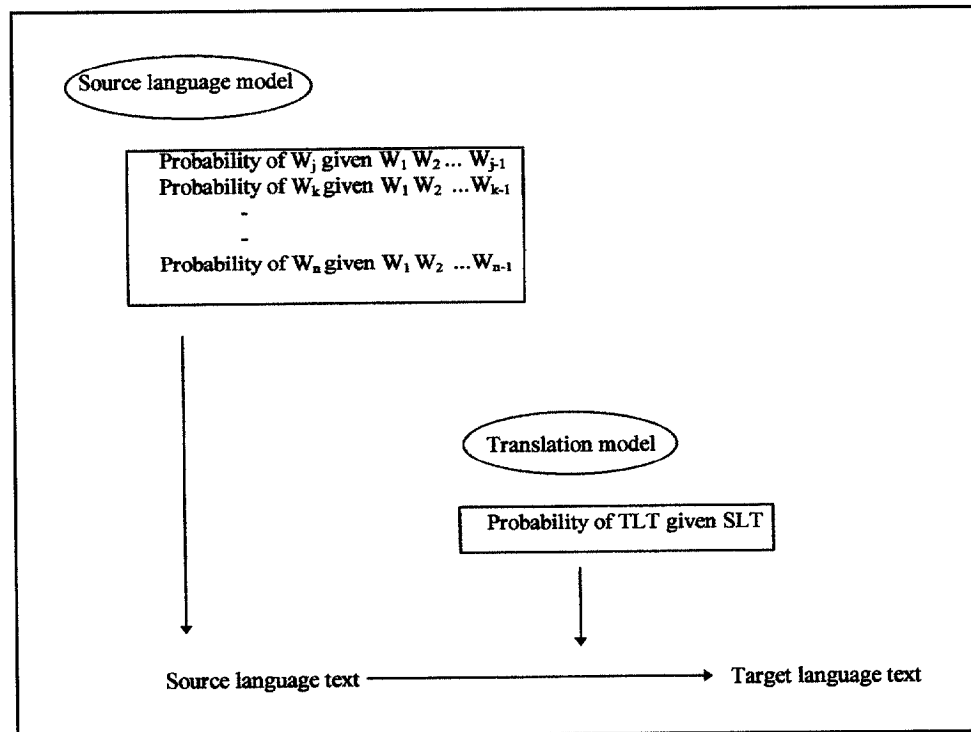


Figure 2:
Translation process in statistics-based MT

CONCLUSIONS

We will not enter here into the controversial issue of rationalist vs. empiricist methods. We believe these two approaches have received adequate attention along the history of science and there is no need to add new arguments to the topic. Our intention here is to point out a new trend in machine translation, one which, in our opinion, seems to be closer to human translation than previous approaches.

Humans translate situations and contexts, not word equivalencies, because words are meaningful only when in use. Very seldom do words have full sense without a context, and this is something we have to consider carefully when our task is translation. Take, for instance, all the possible meanings for the word *leg*, which may require different translations for different contexts. And this is only for just one word, so imagine the importance of context in the translation of a whole text.

According to this, we see corpus-based MT closer to the human activity than “traditional” approaches because, in writing translation rules as contextual equivalencies directly extracted from actual translations, corpus strategies keep words next to the circumstances that help to make up their meanings. On the other hand, “traditional” systems have concentrated on linguistic features of words, usually leaving aside any reference to contextual aspects or finding difficulties in formalising them. We acknowledge the importance of linguistics in building MT systems and, at the same time, emphasize the role of context, which, in our opinion is more easily formalized through corpus strategies.

REFERENCES

- AIJMER, K. and B. ALTERBERG (1991): *English Corpus Linguistics*, London, Longman.
- ALMUALLIM, H., AKIBA, Y., YAMAZAKI, T., YOKOO, A. and S. KANEDA (1994): “Two Methods for Learning ALT-J/E Translation Rules from Examples and a Semantic Hierarchy”, *Proceedings of COLING 94*, pp. 57-63.
- BROWN, P. F. *et al.* (1990): “A Statistical Approach to Machine Translation”, *Computational Linguistics*, 16 (2), pp. 79-85.
- BROWN, P. F., LAIN, J. C. and R. L. MERCER (1991): “Aligning Sentences in Parallel Corpora”, *Proceedings 47th Annual Meeting of the ACL*, pp. 196-176.
- BROWN, P. F. *et al.* (1988): “A Statistical Approach to Language Translation”, *Proceedings of COLING 88*, pp. 71-76.
- BUTLER, C. S. (Ed.) (1992): *Computers and Written Texts*, Oxford, Blackwell.
- CATIZONE, R., RUSSELL, R. and S. WARWICK (in press): “Deriving Translation Data from Bilingual Texts”, *Lexical Acquisition Using On-line Resources to Build a Lexicon*.
- CRANIAS, L., PAPAGEORGIOU, H. and S. PIPERIDIS (1994): “A Matching Technique in Example-based Machine Translation”, *Proceedings of the ACL*, pp. 100-104.
- GALE, W. A. and K. W. CHURCH (1993): “A Program for Aligning Sentences in Bilingual Corpora”, *Computational Linguistics*, 19 (1), pp. 75-102.
- GARSDIE, R., LEECH, G. and G. SAMPSON (Eds.) (1987): *The Computational Analysis of English. A Corpus Based Approach*, London, Longman.
- GOLDFARB, C. F. (1990): *The SGML Handbook*, Oxford, Clarendon Press.
- KAJI, H., KIDA, Y. and Y. MOROMOTO (1992): “Learning Translation Templates from Bilingual Text”, *Proceedings of COLING 92*, pp. 672-678.
- KAY, M. and M. RÖSCHEISEN (1993): “Text Translation Alignment”, *Computational Linguistics*, 19 (1), pp. 121-142.
- KYTO, M., IHALAINEN, O. and M. RISSANEN (Eds.) (1988): *Corpus Linguistics, Hard and Soft. Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora*, Amsterdam, Rodopi.
- LANGEDOEN, TERENCE, D. and F. EANASS (1991): “Feature-structure Markup for Presentation at Oxford and Brown Workshops”, Document number TEI AI W9.
- LANGEDOEN, TERENCE, D., EANASS, F. and G. F. SIMMONS (1993): “A Rationale for the TEI Recommendations for Feature-Structure Markup”, Paper submitted for the Computers and the Humanities special issue on the Text Encoding Initiative, Draft of 22 March 1993, Revised 6 June 1993.
- LEECH, G. (1991): “The State of the Art in Corpus Linguistics”, K. Aijmer and B. Altenberg (Eds.), *English Corpus Linguistics*, London, Longman.

- LEECH, G. (1992): "Corpus Annotation Schemes", *Workshop on Textual Corpora*, Pisa, 24-26 January 1992.
- LEECH, G. and A. WILSON (1994): MSAL 21 "Draft Sections 4.6. and 4.7. of the EAGLES Interim Report: Annotation Sub-Group".
- MEIJS, W. (1987): *Corpus linguistics and Beyond. Proceedings of the Seventh International Conference on English Language Research on Computerized Corpora*, Amsterdam, Rodopi.
- MONACHINI, M. and A. OSTLING (1992a): "Morphosyntactic Corpus Annotation — A Comparison of Different Schemes", *Technical Report*, ILC Pisa, September 1992, NERC-WP8-61.
- MONACHINI, M. and A. OSTLING (1992b): "Towards a Minimal Standard for Morphosyntactic Annotation", *Technical Report*, ILC Pisa, October 1992, NERC-WP8.2.
- MONTEMAGNI, S. (1993): "Syntactically Annotated Corpora: Comparing the Underlying Annotation Schemes", Istituto di Linguistica Computazionale, CNR, Pisa.
- NAGAO, M. (1984): "A Framework of Mechanical Translation between Japanese and English by Analogy Principle", A. Elithorn and R. Banerji (Eds.), *Artificial and Human Intelligence*, Elsevier Science Publishers.
- NIRENBURG, S. *et al.* (1993): "Two Approaches to Matching in Example-based Machine Translation", *Proceedings of TMI-93*, pp. 125-128.
- SADLER, L. and D. ARNOLD (1992): "Unification and Machine Translation", *Meta*, 37 (4), pp. 657-680.
- SADLER, V. and R. VENDELMANS (1990): "Pilot Implementation of a Bilingual Knowledge Bank", *Proceedings of COLING 90*, pp. 449-451.
- SATO, S. and M. NAGAO (1990): "Toward Memory-based Translation", *Proceedings of COLING 90*, pp. 247-252.
- SIMONS, G. F. (1991): "Feature System Declarations and the Interpretation of Feature Structures", Document number TEI AI 1W3 (Draft of 28 January 1991).
- SOUTER, C. and E. ATWELL (Eds.) (1993): *Corpus Based Computational Linguistics*, Amsterdam, Rodopi.
- SPERBERG-MCQUEEN, C. M. and L. BURNARD (Eds.) (1990): *Guidelines for the Encoding and Interchange of Machine-Readable Texts*, Draft version 1.0., Chicago & Oxford, Association for Computers and the Humanities / Association for Computational Linguistics / Association for Literary and Linguistic Computing.
- SUMITA, E. and H. IIDA (1991): "Experiments and Prospects of Example-based Machine Translation", *Proceedings of ACL*, pp. 185-192.
- WEAVER, W. (1949): "Translation", *Machine Translation of Languages*, MIT Press, Cambridge (MA), 1955.
- ZAMPOLLI, A. (1992): "Survey of European Corpus Resources", Presented to UK SALT Club, Wadham College, Oxford, 3 January 1990.
- ZAMPOLLI, A. (1992): "Survey of European Corpus Resources", *Workshop on Textual Corpora*, Pisa, 24-26 January 1992.