

# Corpora and Machine Translation

Harold Somers

Chapter to appear in A. Lüdeling, M. Kytö and T. McEnery (eds) *Corpus Linguistics: An International Handbook*, Berlin, Mouton de Gruyter

## 5 1. Introduction

This chapter concerns the use of corpora in Machine Translation (MT), and, to a lesser extent, the contribution of corpus linguistics to MT and vice versa. MT is of course perhaps the oldest non-numeric application of computers, and certainly one of the first applications of what later became known as natural language processing. However, the early history of MT is marked at first (between roughly 1948 and the early 1960s) by fairly ad hoc approaches as dictated by the relatively unsophisticated computers available, and the minimal impact of linguistic theory. Then, with the emergence of more formal approaches to linguistics, MT warmly embraced – if not exactly a Chomskyan approach – the use of linguistic rule-based approaches which owed a lot to transformational generative grammar. Before this, Gil King (1956) proposed some “stochastic” methods for MT, foreseeing the use of collocation information to help in word-sense disambiguation, and suggesting that distribution statistics should be collected so that, lacking any other information, the most common translation of an ambiguous word should be output (of course he did not use these terms). Such ideas did not resurface for a further 30 years however.

In parallel with the history of corpus linguistics, little reference is made to “corpora” in the MT literature until the 1990s, except in the fairly informal sense of “a collection of texts”. So for example, researchers at the TAUM group (Traduction Automatique Université de Montréal) developed their notion of sublanguage-based MT on the idea that a sublanguage might be defined with reference to a “corpus”: “Researchers at TAUM [...] have made a detailed study of the properties of texts consisting of instructions for aircraft maintenance. The study was based on *a corpus of 70,000 words* of running text in English” (Lehrberger 1982, 207; emphasis added). And in the Eurotra MT project (1983–1990), involving 15 or more groups working more or less independently, a multilingual parallel text in all (at the time) nine languages of the European Communities was used as a “reference corpus” to delimit lexical and grammatical coverage of system. Apart from this, developers of MT systems worked in a largely theory-driven (rather than data-driven) manner, as characterised by Isabelle (1992a) in his Preface to the Proceedings of the landmark TMI Conference of that year: “On the one hand, the “rationalist” methodology, which has dominated MT for several decades, stresses the importance of basing MT development on better theories of natural language.... On the other hand, there has been renewed interest recently in more “empirical” methods, which give priority to the analysis of large corpora of existing translations....”

The link between MT and corpora really first became established with the emergence of statistics-based MT (SMT) from 1988 onwards. The IBM group at Yorktown Heights, NY had got the idea of doing SMT based on their success with speech recognition, and then had to look round for a suitable corpus (Fred Jelinek, personal communication). Fortunately, the Canadian parliament had in 1986 started to make its bilingual (English and French) proceedings (Hansard) available in machine-readable form. However, again, the “corpus” in question was really just a collection of raw text, and the MT methodology had no need in the first instance of any sort of mark-up or annotation (cf. Articles 20 and 34). In Section 4 we

45 will explain how SMT works and how it uses techniques of interest to corpus linguists.

The availability of large-scale parallel texts gave rise to a number of developments in the MT world, notably the emergence of various tools for translators based on them, the “translation memory” (TM) being the one that has had the greatest impact, though parallel concordancing also promises to be of great benefit to translators (see Sections 1.1 and 1.2). Both of these applications rely on the parallel text having been *aligned*, techniques for which are described  
50 in Articles 20 and 34. Not all TMs are corpus-based however, as will be discussed in Section 1.2.

Related to, but significantly different from TMs, is an approach to MT termed “Example-Based MT” (EBMT). Like TMs, this takes the idea that new translations can use existing  
55 translations as a model, the difference being that in EBMT it is the computer rather than the translator that decides how to manipulate the existing example. As with TMs, not all EBMT systems are corpus-based, and indeed the provenance of the examples that are used to populate the TM or the example-base is an aspect of the approach that is open to discussion. Early EBMT systems tended to use hand-picked examples, whereas the latest developments in  
60 EBMT tend to be based more explicitly on the use of naturally occurring parallel corpora also making use in some cases of mark-up and annotations, this extending in one particular approach, to tree banks (cf. Articles 17 and 29). All these issues are discussed in Section 3. Recent developments in EBMT and SMT have seen the two paradigms coming closer together, to such an extent that some commentators doubt there is a significant difference.

65 One activity that sees particular advantage in corpus-based approaches to MT, whether SMT or EBMT, is the rapid development of MT for less-studied (or “low density”) languages (cf. Article 23). The essential element of corpus-based approaches to MT is that they allow systems to be developed automatically, in theory without the involvement of language experts or native speakers. The MT systems are built by programs which “learn” the translation  
70 relationships from pre-existing translated texts, or apply methods of “analogical processing” to infer new translations from old. This learning process may be helped by some linguistically-aware input (for example, it may be useful to know what sort of linguistic features characterise the language pair in question) but in essence the idea is that an MT system for a new language pair can be built just on the basis of (a sufficient amount of)  
75 parallel text. This is of course very attractive for “minority” languages where typically parallel texts such as legislation or community information in both the major and minor languages exists. Most of the work in this area has been using the SMT model, and we discuss these developments in Section 6.

## 2. Corpus-based tools for translators

80 Since the mid-1980s, parallel texts in (usually) two languages have become increasingly available in machine-readable form. Probably the first such “bitext” of significant size, to use the term coined by Harris (1988), was the Canadian Hansard mentioned above. The Hong Kong parliament, with proceedings at that time in English and Cantonese, soon followed suit, and the parallel multilingual proceedings of the European Parliament are a rich source of data;  
85 but with the explosion of the World Wide Web, parallel texts, sometimes in several languages, and of varying size and quality, soon became easily available.

Isabelle (1992b, 8) stated that “*Existing translations contain more solutions to more translation problems than any other existing resource*” [emphasis original], reflecting the idea,

90 first proposed independently by Arthern (1978), Kay (1980) and Melby (1981), that a store of past translations together with software to access it could be a useful tool for translators. The realisation of this idea had to wait some 15 years for adequate technology, but is now found in two main forms, parallel concordances, and TMs.

### 1.1 *Parallel concordances*

95 Parallel concordances have been proposed for use by translators and language learners, as well as for comparative linguistics and literary studies where translation is an issue (e.g. with biblical and quranic texts). An early implementation is reported by Church and Gale (1991), who suggest that parallel concordancing can be of interest to lexicographers, illustrated by the ability of a parallel concordance to separate the two French translations of *drug* (*médicament* ‘medical drug’ vs. *drogue* ‘narcotic’). An implementation specifically aimed at translators is  
100 *TransSearch*, developed since 1993 by RALI in Montreal (Simard et al. 1993), initially using the Canadian Hansard, but now available with other parallel texts. Part of a suite of *Trans-*tools, *TransSearch* was always thought of as a translation aid, unlike *ParaConc* (Barlow 1995) which was designed for the purpose of comparative linguistic study of translations, and *MultiConcord* (Romary et al. 1995), aimed at language teachers. More recently, many articles  
105 dealing with various language combinations have appeared. In each case, the idea is that one can search for a word or phrase in one language, and retrieve examples of its use in the normal manner of a (monolingual) concordance, but in this case linked (usually on a sentence-by-sentence basis) to their translations. Apart from its use as a kind of lexical look-up, the concordance can also show contexts which might help differentiate the usage of alternate  
110 translations or near synonyms. Most systems also allow the use of wildcards, but also parallel search, so that the user can retrieve examples of a given source phrase coupled with a target word. This device can be used, among other things, to check for false-friend translations (e.g. French *librairie* as *library* rather than *bookshop*), or to distinguish, as above, different word senses.

115 A further use of a parallel corpus as a translator’s aid is the RALI group’s *TransType* (Foster et al. 2002), which offers translators text completion on the basis of the parallel corpus. With the source text open in one window, the translator starts typing the translation, and on the basis of the first few characters typed, the system tries to predict from the target-language side of the corpus what the translator wants to type. This predication capability is enhanced by  
120 Maximum Entropy, word- and phrase-based models of the target language and some techniques from Machine Learning. Part of the functionality of *TransType* is like a sophisticated TM, the increasingly popular translator’s aid that we will discuss in the next section.

### 1.2 *Translation Memories (TMs)*

125 The TM is one of the most significant computer-based aids for translators. First proposed independently by Arthern (1978), Kay (1980) and Melby (1981), but not generally available until the mid 1990s (see Somers and Fernández Díaz, 2004, 6–8 for more detailed history), the idea is that the translator can consult a database of previous translations, usually on a sentence-by-sentence basis, looking for anything similar enough to the current sentence to be  
130 translated, and can then use the retrieved example as a model. If an exact match is found, it can be simply cut and pasted into the target text, assuming the context is similar. Otherwise, the translator can use it as a suggestion for how the new sentence should be translated. The TM will highlight the parts of the example(s) that differ from the given sentence, but it is up

to the translator to decide which parts of the target text need to be changed.

135 One of the issues for TM systems is where the examples come from: originally, it was thought that translators would build up their TMs by storing their translations as they went along. More recently, it has been recognised that a pre-existing bilingual parallel text could be used as a ready-made TM, and many TM systems now include software for aligning such data (see Article 20).

140 Although a TM is not necessarily a “corpus”, strictly speaking, it may still be of interest to discuss briefly how TMs work and what their benefits and limitations are. For a more detailed discussion, see Somers (2003).

### 1.2.1 Matching and equivalence

145 Apart from the question of where the data comes from, the main issue for TM systems is the problem of matching the text to be translated against the database so as to extract all and only the most useful cases to help and guide the translator. Most current commercial TM systems offer a quantitative evaluation of the match in the form of a “score”, often expressed as a percentage, and sometimes called a “fuzzy match score” or similar. How this score is arrived at can be quite complex, and is not usually made explicit in commercial systems, for  
150 proprietary reasons. In all systems, matching is essentially based on character-string similarity, but many systems allow the user to indicate weightings for other factors, such as the source of the example, formatting differences, and even significance of certain words. Particularly important in this respect are strings referred to as “placeables” (Bowker 2002, 98), “transwords” (Gaussier et al. 1992, 121), “named entities” (using the term found in  
155 information extraction) Macklovitch and Russell 2000, 143), or, more transparently perhaps, “non-translatables” (ibid., 138), i.e. strings which remain unchanged in translation, especially alphanumerics and proper names: where these are the only difference between the sentence to be translated and the matched example, translation can be done automatically. The character-string similarity calculation uses the well-established concept of “sequence comparison”, also  
160 known as the “string-edit distance” because of its use in spell-checkers, or more formally the “Levenshtein distance” after the Russian mathematician who discovered the most efficient way to calculate it. A drawback with this simplistic string-edit distance is that it does not take other factors into account. For example, consider the four sentences in (1).

- 165 (1) a. Select ‘Symbol’ in the Insert menu.  
b. Select ‘Symbol’ in the Insert menu to enter a character from the symbol set.  
c. Select ‘Paste’ in the Edit menu.  
d. Select ‘Paste’ in the Edit menu to enter some text from the clip board.

170 Given (1a) as input, most character-based similarity metrics would choose (1c) as the best match, since it differs in only two words, whereas (1b) has eight additional words. But intuitively (1b) is a better match since it entirely includes the text of (1a). Furthermore (1b) and (1d) are more similar than (1a) and (1c): the latter pair may have fewer words different (2 vs. 6), but the former pair have more words in common (8 vs. 4), so the distance measure should count not only differences but also similarities.

175 The similarity measure in the TM system may be based on individual characters or whole words, or may take both into consideration. Although more sophisticated methods of matching have been suggested, incorporating linguistic “knowledge” of inflection paradigms,

synonyms and even grammatical alternations (Cranias et al. 1997, Planas and Furuse 1999, Macklovitch and Russell 2000, Rapp 2002), it is unclear whether any existing commercial systems go this far. To exemplify, consider (2a). The example (2b) differs only in a few characters, and would be picked up by any currently available TM matcher. (2c) is superficially quite dissimilar, but is made up of words which are related to the words in (2a) either as grammatical alternatives or near synonyms. (2d) is very similar in meaning to (2a), but quite different in structure. Arguably, any of (2b–d) should be picked up by a sophisticated TM matcher, but it is unlikely that any commercial TM system would have this capability.

- (2) a. When the paper tray is empty, remove it and refill it with paper of the appropriate size.  
 b. When the tray is empty, remove it and fill it with the appropriate paper.  
 c. When the bulb remains unlit, remove it and replace it with a new bulb  
 d. You have to remove the paper tray in order to refill it when it is empty.

The reason for this is that the matcher uses a quite generic algorithm, as mentioned above. If we wanted it to make more sophisticated *linguistically*-motivated distinctions, the matcher would have to have some language-specific “knowledge”, and would therefore have to be different for different languages. It is doubtful whether the gain in accuracy would merit the extra effort required by the developers. As it stands, TM systems remain largely independent of the source language and of course wholly independent of the target language.

Nearly all TM systems work exclusively at the level of sentence matching. But consider the case where an input such as (3) results in matches like those in (4).

- (3) Select ‘Symbol’ in the Insert menu to enter a character from the symbol set.  
 (4) a. Select ‘Paste’ in the Edit menu.  
 b. To enter a symbol character, choose the Insert menu and select ‘Symbol’.

Neither match covers the input sentence sufficiently, but between them they contain the answer. It would clearly be of great help to the translator if TM systems could present partial matches and allow the user to cut and paste fragments from each of the matches. This is being worked on by most of the companies offering TM products, and, in a simplified form, is currently offered by at least one of them, but in practice works only in a limited way, for example requiring the fragments to be of roughly equal length (see Somers & Fernández Díaz 2004).

### 1.2.2 Suitability of naturally occurring text

As mentioned above, there are two possible sources of the examples in the TM database: either it can be built up by the user (called “interactive translation” by Bowker 2002, 108), or else a naturally occurring parallel text can be aligned and used as a TM (“post-translation alignment”, *ibid.*, 109). Both methods are of relevance to corpus linguists, although the former only in the sense that a TM collected in this way could be seen as a special case of a planned corpus. The latter method is certainly quicker, though not necessarily straightforward (cf. Macdonald 2001), but has a number of shortcomings, since a naturally occurring parallel text will not necessarily function optimally as a TM database.

The first problem is that it may contain repetitions, so that a given input sentence may

220 apparently have multiple matches, but they might turn out to be the same. This of course could be turned into a good thing, if the software could recognize that the same phrase was being consistently translated in the same way, and this could bolster any kind of confidence score that the system might calculate for the different matches.

225 More likely though is that naturally occurring parallel text will be internally *inconsistent*: a given phrase may have multiple translations either because different translations are appropriate in different contexts, or because the phrase has been translated in different ways for no reason other than that translators have different ideas or like to introduce variety into their translations. Where different contexts call for different translations, then the parallel corpus is of value assuming that it can show the different contexts, as discussed in the previous section. For example, the simple phrase *OK* in a conversation may be translated into  
 230 Japanese as *wakarimashita* ‘I understand’, *iidesu yo* ‘I agree’ or *ijō desu* ‘let’s change the subject’, depending on the context (example from Somers et al. 1990, 403). However, this is not such a big problem because the TM is a translator’s *tool*, and in the end the responsibility for choosing the translation is the user’s. The problem of suitability of examples is more serious in EBMT, as we will discuss below.

### 235 3. Example-based MT (EBMT)

EBMT is often thought of as a sophisticated type of TM, although in fact this approach to MT initially developed somewhat independently of the TM idea, albeit around the same time. In this section we will explain briefly how it works, and clarify some important differences between TMs and EBMT.

240 The idea for EBMT surfaced in the early 1980s (the seminal paper presented by Makoto Nagao at a 1981 conference was not published until three years later – Nagao, 1984), but the main developments were reported from about 1990 onwards, and it has slowly become established within the mainstream of MT research (cf. Carl and Way 2003, 2006/7). Pioneers were mainly in Japan, including Sato and Nagao (1990) and Sumita et al. (1990). As in a TM,  
 245 the basic idea is to use a database of previous translations, the “example-base”, and the essential first step, given a piece of text to translate, is to find the best match(es) for that text. Much of what was said above regarding matching in TM systems also applies to EBMT, though it should be said that earlier implementations of EBMT often had much more complex matching procedures, linked to the fact that examples were often stored not just as plain text  
 250 but as annotated tree or other structures, often explicitly aligned.

Once the match has been found, the two techniques begin to diverge. While the work of the TM system is over (the translator decides what to do with the matches), in EBMT the system now has to manipulate the example so as to produce a translation. This is done in three steps:  
 255 first, the source text and the examples are aligned so as to highlight which parts of the examples correspond to text in the sentence to be translated. Next, and crucially, the corresponding target-language fragments of text must be identified in the translations associated with the matches. Finally, the target translation is composed from the fragments so identified.

260 We can illustrate the process with a simple example. Suppose the input sentence is (5), and the matching algorithm identifies as relevant to its translation the examples in (6) with their French translations. The fragments of text in the examples that match the input are underlined.

- (5) The operation was interrupted because the file was hidden.  
 (6) a. The operation was interrupted because the Ctrl-c key was pressed.  
       *L'opération a été interrompue car la touché Ctrl-c a été enfoncée.*  
 265       b. The specified method failed because the file is hidden.  
           *La méthode spécifiée a échoué car le fichier est masqué.*

The EBMT process must now pick out from the French examples in (6) which words correspond to the underlined English words, and then combine them to give the proposed translation. These two operations are known in the EBMT literature as “alignment” and “recombination”.

### 1.3 Alignment in EBMT

Alignment, similar to but not to be confused with the notion of aligning parallel corpora in general, involves identifying which words in the French sentences correspond to the English words that we have identified as being of interest. An obvious way to do this might be with the help of a bilingual dictionary, and indeed some EBMT systems do work this way. However, one of the attractions of EBMT is the idea that an MT system can be built up on the basis only of large amounts of parallel data, with lexical alignments extracted from the examples automatically by analogy. This idea is of interest to corpus linguists, and indeed there is a literature around this topic (cf. Article 34). In particular, techniques relying on simple probabilities using contingency tables and measures such as Dice’s coefficient, are well explored.

Within EBMT, there is a strand of research which seeks to generalize similar examples and thereby extract lexical correspondences, as follows: suppose that in the example base we have the sentences in (7), with their corresponding Japanese translations.

- 285       (7) a. The monkey ate a peach. ↔ *Saru wa momo o tabeta.*  
           b. The man ate a peach. ↔ *Hito wa momo o tabeta.*

From the sentence pairs in (7) we can assume that the difference between the two English sentences, *monkey* vs. *man*, corresponds to the only difference between the two Japanese translations, *saru* vs. *hito*. Furthermore we can assume that the remaining parts which the two sentences have in common also represent a partial translation pair (8).

- (8) The X ate a peach. ↔ *X' wa momo o tabeta.*

Comparison with further examples which are minimally different will allow us to build up both a lexicon of individual word pairs, and a “grammar” of transfer template pairs. Ideas along these lines have been explored for example by Cicekli and Güvenir (1996), Cicekli (2006), Brown (2000, 2001) and by McTait and Trujillo (1999).

### 1.4 Recombination

Once the appropriate target-language words and fragments have been identified, it should be just a matter of sticking them together. At this stage however a further problem arises, generally referred to in the literature as “boundary friction” (Nirenburg et al. 1993, 48; Collins 1998, 22): fragments taken from one context may not fit neatly into another slightly different context. For example, if we have the translation pair in (9) and replace *man* with *woman*, the

resulting translation, with *homme* replaced by *femme* is quite ungrammatical, because French requires gender agreement between the determiner, adjective and noun.

(9) The old man is dead.  $\leftrightarrow$  *Le vieil homme est mort.*

305 Another problem is that the fragments to be pasted together sometimes overlap: if we look again at examples (5) and (6), the fragments we have to recombine are the French equivalents of the templates shown in (10a,b) from (6a,b) respectively.

(10) a. The operation was interrupted because the ... was ....  
b. The ... because the file ... hidden.

310 A number of solutions to these two difficulties have been suggested, including the incorporation of target-language grammatical information which itself might be derived from a parallel corpus (Wu 1997), or, of more interest to corpus linguists, a model of target-language word sequences, or matching the proposed target sentence against the target side of the bilingual corpus.

#### 315 **4. Statistical MT (SMT)**

SMT in its various forms is probably the approach to MT whose techniques and methodologies are most familiar to corpus linguists. In this section, we will discuss briefly the main ideas behind SMT, and some of the latest developments.

320 In its pure form, the statistics-based approach to MT makes use of no traditional linguistic data. The essence of the method is first to align phrases, word groups and individual words of the parallel texts, and then to calculate the probabilities that any one word in a sentence of one language corresponds to a word or words in the translated sentence with which it is aligned in the other language. An essential feature is the availability of a suitable large bilingual corpus of reliable (authoritative) translations.

325 The “empirical” approach to MT was pioneered by the IBM research group at Yorktown Heights, NY, who had had some success with non-linguistic approaches to speech recognition, and turned their attention to MT in the early 1990s (Brown et al. 1990). Perhaps because of the promise this approach showed – systems could be built in a matter of weeks which came fairly close to the quality of rule-based systems which had taken many person-  
330 years to build – or simply because of the attraction of a rather new slant on an old problem, an SMT approach was taken up by a number of groups.

##### *1.5 How it works*

As already mentioned, the idea is to “model” the translation process in terms of statistical probabilities. For a given source-language sentence  $S$ , there are an infinite number of  
335 “translations”  $T$  of varying probability. The idea of SMT is to find just the  $T$  that maximizes the probability  $P(T|S)$ . This probability is seen as a function of two elements: a set  $\{t_1, t_2, \dots, t_m\}$  of most probable target-language words given the set of source-language words  $\{s_1, s_2, \dots, s_n\}$  which make up  $S$ , and the most probable order in which that given set of target-language words might occur. These two elements are referred to as the “translation model” and the  
340 “language model” respectively. Both are computed on the basis of the bilingual corpus.

The translation process in SMT therefore consists of applying the translation model to a given source sentence  $S$  to produce a set of probable words, and then applying the language model to those words to produce the target sentence  $T$ . However, since there are different probabilities involved, this is not a straightforward calculation, because the different probabilities interact. In effect, we start with the target-language words which look most likely to be part of the solution, see how these choices fit with the language model, and, in a systematic way, keep trying different combinations until we cannot improve the overall “score” any more. This so-called “decoding” stage of SMT is further discussed below.

### 1.5.1 The translation model

The translation model is the set of probabilities for each word on the source-language side of the corpus that it corresponds to or gives rise to each word on the target-language side of the corpus. Of course for many of these word pairs, the probability will be close to 0. The hope is that for words which are translational equivalents, the probabilities will be suitably high. One problem for this approach is that, as all linguists know, there is generally not a 1:1 correspondence between the words of one language and another. For example, French translations of adjectives in English have different forms depending on gender and number agreement. Homonyms in one language will have different translations in the target language. Importantly also some single words in one language may be translated by a string of words in the other language: for example, the single word *implemented* in English may be rendered in French as *mise en application*. This is referred to as the “fertility” of the source-language word. For this reason, the translation model includes not just word-pair translation probabilities, but a second set of parameters measuring the probability of different fertilities.

For practical reasons, these may be restricted to a small given range, for example 0–2 (0, because a word on the source side may “disappear” in translation, for example the two English words *may have* give rise to just one French word *aurait*). Fertility is nicely illustrated in the original IBM work (Brown et al. 1990), with data from the bilingual Canadian Hansards. The English word *the* translates as *le* with  $P=.610$ , *la* with  $P=.178$ , and some other words with much smaller values. Fertility  $f=1$  with a .817 probability. The word *not* on the other hand translates as *pas* with  $P=.469$ , *ne* with  $P=.460$ , that is, with roughly equal probability. The fertility probabilities are .758 for  $f=2$ , .133 for  $f=0$  and .106 for  $f=1$ . In other words, the French for *not* is very likely to be *ne...pas*. One last example is particular to the Hansard corpus. Very frequent in this corpus is the English phrase *hear hear*. The English word *hear* is coupled with the French *bravo* with  $P=.992$  (and with much lower probabilities to various forms of the French verb *entendre*); the fertility probabilities are almost evenly split between  $f=0$  ( $P=.584$ ) and  $f=1$  ( $P=.416$ ). In other words, *hear* is almost certain to be translated as *bravo*, when it is translated at all, but half the time it should be simply omitted.

One can imagine various different methods of computing these probabilities. Assuming that the bilingual corpus is sentence-aligned, and based on their experience with speech recognition, Brown et al. (1990) use the Expectation-Maximization (EM) algorithm to compute the most likely word alignments, allowing only 1:0, 1:1 and 1:2 couplings (notably not 0:n, or many:many).

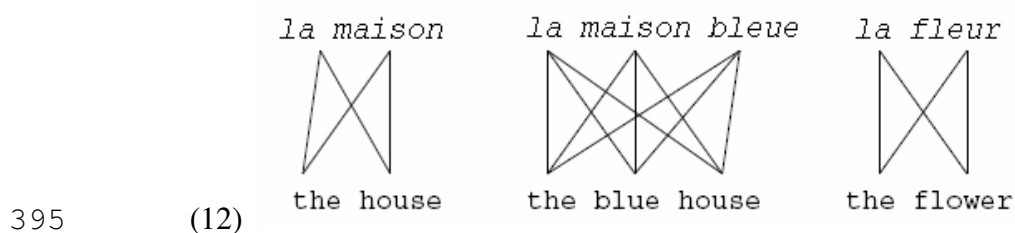
### 1.5.2 Word alignment with the EM algorithm

The EM algorithm (Dempster et al. 1977) is widely used in a variety of tasks involving incomplete data, where missing parameters are estimated, then these estimates are used to

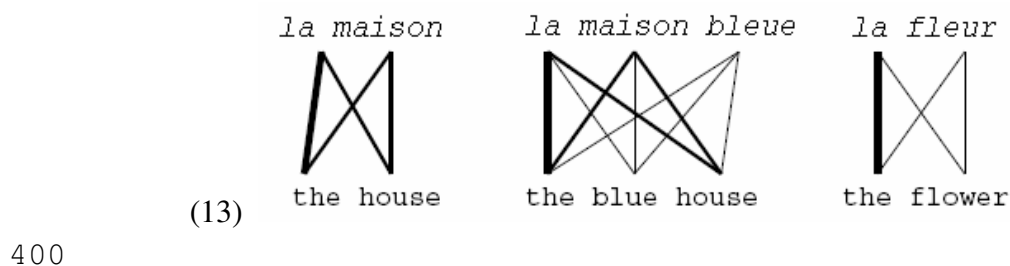
385 retrain the model, the process iterating until the best results are obtained. We can illustrate its  
 use in word alignment for translation modelling by considering the examples in (11) (from  
 Knight and Koehn 2004), which we assume to have been extracted from a larger corpus  
 containing many other translation pairs.

390 (11) *la maison* ↔ the house  
*la maison bleue* ↔ the blue house  
*la fleur* ↔ the flower

Initially, we assume that all word alignments are equally likely, as in (12), but a first pass  
 shows that 3 of the 17 connections link *la* with *the*, and 2 out of 17 link with *la* with *house*,  
*maison* with *house*, and *maison* with *the*.



A first iteration strengthens these more likely connections and at the same time weakens  
 connections that are in conflict with them (13).



Further iterations strengthen connections such as the one between *fleur* and *flower*: because  
*la* is linked with *the*, *flower* is the only link open for *fleur* (14).



405 Eventually, there is convergence, and the inherent structure (15) is arrived at.



Obviously, the perfect alignment as seen in (15) is an ideal result: in practice, the resulting alignments are a set of probabilities, which reflect the alignments over the corpus. For example, besides the alignment of *la* with *the*, one could expect in a corpus that there would also be evidence for aligning *le* and *les* with *the*, and probabilities would reflect the relative strengths of these pieces of evidence.

### 1.5.3 The IBM Models

Brown et al. (1993) suggested a number of different ways in which their original (1990) basic model could be enhanced, in what have become known as “IBM Models” 1–5. In what follows we give a necessarily brief overview of the five models: for mathematical details readers are referred to the original source. The simplest, Model 1, assumes a uniform distribution, i.e. that the target-language word should occur in the place in the sequence corresponding to its place in the source-language sequence. Model 2 tries to model relative position in the word stream by calculating the probabilities of a certain position in the target string for each word given its position in the source string, and the lengths of the two strings: a word near the beginning of the source sentence is more likely to correspond to a word near the beginning of the target sentence, especially if the sentence is long. Model 3 includes fertility probabilities, as described above, and modelling distortion better. Model 4 additionally takes into account the fact that often words in the source language constitute a phrase which is translated as a unit in the target language. For example, in the translation pair in (16), *nodding* is associated with the phrase *faire signe que oui* in Model 4, while in Model 3 it is connected only to the words *signe* and *oui*.

(16) *Il me semble faire signe que oui.*  
It seems to me that he is nodding.

Finally, Model 5 rectifies a deficiency in Models 3 and 4 whereby words can be assigned to the same position in the target string, or to positions before or after the start or end of the target string.

Other researchers have typically taken one of models 1–3 as a starting point, and tried to develop strategies from there (see Och and Ney 2003).

Some alternatives to the word-based IBM models have emerged more recently: we will discuss these approaches in Section 5.

### 1.5.4 The target language model

As mentioned above, the aim of the language model is to predict the most likely sequence of target-language words chosen by the translation model. To some extent, word-sequence is determined by the translation model (especially the higher-numbered IBM models, and also more recent approaches to be discussed in Section 5), but the language model is necessary to

ensure that target-language-specific features, such as agreement and long-distance dependencies, not easily captured by the translation model, are covered.

445 The target-language model essentially models the probability of sequences of words. In principle, we could model the probability of a sequence of words  $w_1, w_2, \dots, w_m$ , by modelling the probability of each successive word given the preceding sequence  $P(w_i | w_1, \dots, w_{i-1})$ , so that the probability of the entire sequence would be (17).

$$(17) P(w_1, \dots, w_m) = \prod_{i=1}^m P(w_i | w_1, \dots, w_{i-1})$$

450 Unfortunately, this is impractical or even impossible, given the infinite nature of language and the problem of sparse data (see [Article 39](#)). In practice, it turns out that the trade-off between practicality and usability comes if we look only at sequences of 3 or 4 words, referred to as  $n$ -grams with  $n=3$  or  $n=4$ . The probability of a given string of words using a trigram model is given by (18).

$$(18) P(w_1, \dots, w_m) = \prod_{i=1}^m P(w_i | w_{i-2}, w_{i-1})$$

455 Probabilities for  $i < 3$  are catered for by considering start-of-sentence as a pseudo-word. One problem with this model is again that of sparse data: if any of the trigrams happen not to occur in the training data, as is very likely, they will receive a 0 probability score, which will of course result in the product being 0. This is overcome in two ways: smoothing, and back-off. “Smoothing” consists of adjusting all the parameters so that none of them are 0. Crudely, one  
460 could add a tiny value to each parameter, but there are numerous other better motivated methods of smoothing (see Manning and Schütze 1999, 199ff; Jurafsky and Martin 2000, 206ff). “Back off” involves looking at  $(n-1)$ -gram models if the  $n$ -gram is unseen. So a trigram model would back off to include bigram and if necessary unigram statistics, as in (19),

$$(19) \hat{P}(w_i | w_{i-2}, w_{i-1}) = \begin{cases} P(w_i | w_{i-2}, w_{i-1}) & \text{if } f(w_{i-2} w_{i-1} w) > 0 \\ \alpha P(w_i | w_{i-1}) & \text{if } f(w_{i-2} w_{i-1} w) = 0 \\ & \text{and } f(w_{i-1} w) > 0 \\ \beta P(w_i) & \text{otherwise} \end{cases}$$

465 where  $f$  is the frequency count of the  $n$ -gram sequence, and  $\alpha, \beta$  are weights (see Manning and Schütze 1999, 219ff; Jurafsky and Martin 2000, 216ff).

### 1.5.5 The decoder

To complete the SMT system we need a program which can apply the translation and language models to a given input text, that is to search for the target text which maximizes the  
470 probability equations. This part of the process has come to be known as the “decoder”. Evidently, given the number of statistical parameters, an exhaustive search of all possible combinations is impractical. Knight (1999) demonstrated that the problem was NP-complete. Various alternatives have been proposed for this problem.

The simplest is perhaps the “stack search” which basically starts by proposing a simple

475 hypothesis, for example take the most probable word-by-word translation in the order of the  
 source text, and explore the surrounding search space in a motivated way until the “score”  
 cannot be further improved (Wang and Waibel 1997; Germann et al. 2001). A variant of this  
 is a “beam search”. In this case the target text is built up left-to-right by expanding the  
 translation hypotheses. Since this method is exponential in the length of the sentence, various  
 480 tactics are needed to make the search more tractable. Pruning obviously weak hypotheses is a  
 good start, for example eliminating texts where the number of words in the source and target  
 texts are vastly different. Ueffing et al. (2002) used word graphs to maximise the efficiency of  
 the beam search. Decoding viewed as state-space search to be tackled using methods based on  
 Dynamic Programming is an approach taken by Garcia-Varea et al. (1998), Niessen et al.  
 485 (1998), Och et al. (2001) and Tillman and Ney (2003). Tillmann et al. (1997) use an approach  
 based on Hidden Markov Model alignments. Watanabe and Sumita (2003) present a method  
 which uses some techniques borrowed from EBMT.

## 5. Variants of SMT

490 Early on in the history of SMT it was recognised that simple word-based models would only  
 go so far in achieving a reasonable quality of translation. In particular, cases where single  
 words in one language are translated as multi-word phrases in the other, and cases where the  
 target-language syntax is significantly distorted with respect to the source language often  
 cause bad translations in simple SMT models. Examples of these two phenomena are to be  
 found when translating between German and English, as seen in (20)-(21) (from Knight and  
 495 Koehn 2004).

(20) a. *Zeitmangel erschwert das Problem.*

lit. Lack-of-time makes-more-difficult the problem  
 ‘Lack of time makes the problem more difficult.’

b. *Eine Diskussion erübrigt sich demnach.*

500 lit. A discussion makes-unnecessary itself therefore  
 ‘Therefore there is no point in discussion.’

(21) a. *Das ist der Sache nicht angemessen.*

lit. That is to-the matter not appropriate  
 ‘That is not appropriate for this matter.’

505 b. *Den Vorschlag lehnt die Kommission ab.*

lit. The proposal rejects the Commission off  
 ‘The Commission rejects the proposal.’

To address these problems, variations of the SMT model have emerged which try to work  
 with phrases rather than words, and with structure rather than strings. These approaches are  
 510 described in the next two sections.

### 1.5.6 Phrase-based SMT

The idea behind “phrase-based SMT” is to enhance the conditional probabilities seen in the  
 basic models with joint probabilities, i.e. “phrases”. Because the alignment is again purely  
 statistical, the resulting phrases need not necessarily correspond to groupings that a linguist  
 515 would identify as constituents.

Wang and Waibel (1998) proposed an alignment model based on shallow model structures.  
 Since their translation model reordered phrases directly, it achieved higher accuracy for

520 translation between languages with different word orders. Other researchers have explored the idea further (Och et al. 1999, Marcu and Wong 2002, Koehn and Knight 2003, Koehn et al. 2003).

525 Och and Ney's (2004) alignment template approach takes the context of words into account in the translation model, and local changes in word order from source to target language are learned explicitly. The model is described using a log-linear modelling approach, which is a generalization of the often used source-channel approach. This makes the model easier to extend than classical SMT systems. The system has performed well in evaluations.

To illustrate the general idea more exactly, let us consider (22) as an example (from Knight and Koehn 2004).

530 (22) *Maria no daba una bofetada a la bruja verde.*  
lit. Maria not gave a slap to the witch green  
'Maria did not slap the green witch.'

535 First, the word alignments are calculated in the usual way. Then potential phrases are extracted by taking word sequences which line up in both the English and Spanish, as in Figure 1.

|       | Maria | no | daba | una | bofetada | a | la | bruja | verde |
|-------|-------|----|------|-----|----------|---|----|-------|-------|
| Maria |       |    |      |     |          |   |    |       |       |
| did   |       |    |      |     |          |   |    |       |       |
| not   |       |    |      |     |          |   |    |       |       |
| slap  |       |    |      |     |          |   |    |       |       |
| the   |       |    |      |     |          |   |    |       |       |
| green |       |    |      |     |          |   |    |       |       |
| witch |       |    |      |     |          |   |    |       |       |

Figure 1. Initial phrasal alignment for example (22)

If we take all sequences of contiguous alignments, this gives us possible phrase alignments as in (23) for which probabilities can be calculated based on the relative co-occurrence frequency of the pairings in the rest of the corpus.

540 (23) (Maria, *Maria*)  
(did not, *no*)  
(slap, *daba una bofetada*)  
(the, *a la*)  
(green, *verde*)  
545 (witch, *bruja*)

By the same principle, a further iteration can identify larger phrases, as long as the sequences are contiguous, as in Figure 2.

|       | Maria | no | daba | una | bofetada | a | la | bruja | verda |
|-------|-------|----|------|-----|----------|---|----|-------|-------|
| Maria | ■     |    |      |     |          |   |    |       |       |
| did   |       | ■  |      |     |          |   |    |       |       |
| not   |       |    | ■    |     |          |   |    |       |       |
| slap  |       |    |      | ■   | ■        | ■ |    |       |       |
| the   |       |    |      |     |          |   | ■  | ■     |       |
| green |       |    |      |     |          |   |    |       | ■     |
| witch |       |    |      |     |          |   |    | ■     |       |

Figure 2. Further phrasal identification

- (24) (Maria did not, *Maria no*)  
 (did not slap, *no daba una bofetada*)  
 (slap the, *daba una bofetada a la*)  
 (green witch, *bruja verda*)

555 The process continues, each time combining contiguous sequences giving the phrases in (25), (26) and finally (27), the whole sentence.

- (25) (Maria did not slap, *Maria no daba una bofetada*)  
 (did not slap the, *no daba una bofetada a la*)  
 (the green witch, *a la bruja verda*)

560 (26) (Maria did not slap the, *Maria no daba una bofetada a la*)  
 (slap the green witch, *daba una bofetada a la bruja verda*)

- (27) (Maria did not slap the green witch, *Maria no daba una bofetada a la bruja verda*)

Of course, as the phrases get longer, the probabilities get smaller, as their frequency in the corpus diminishes.

565 Koehn et al. (2003) evaluated a number of variants of the phrase-based SMT approach, and found that they all represented an improvement over the original word-based approaches. Furthermore, increased corpus size had a more marked positive effect than it did with word-based models. The best results were obtained when the probabilities for the phrases were weighted to reflect lexical probabilities, i.e. scores for individual word-alignments. And, most  
 570 interestingly, if phrases not corresponding to constituents in a traditional linguistic view were excluded, the results were not as good.

### 1.5.7 Structure-based SMT

575 Despite the improvements, a number of linguistic phenomena still prove troublesome, notably discontinuous phrases and long-distance reordering, as in (21). To try to handle these, the idea of “syntax-based SMT” or “structure-based SMT” has developed, benefiting from ideas from stochastic parsing and the use of treebanks (see Articles 7, 17, 29).

Wu (1997) introduced Inversion Transduction Grammars as a grammar formalism to provide structural descriptions of two languages simultaneously, and thereby a mapping between them: crucially, his grammars of English and Cantonese were derived from the bilingual

580 Hong Kong Hansard corpus. The development of an efficient decoder based on Dynamic Programming permits the formalism to be used for SMT (Wu and Wong 1998). Alshawi et al. (1998) developed a hierarchical transduction model based on finite-state transducers: using an automatically induced dependency structure, an initial head-word pair is chosen, and the sentence is then expanded by translating the dependent structures. In Yamada and Knight's  
 585 (2001) "tree-to-string" model a parser is used on the source text only. The tree is then subject to reordering, insertion and translation operations, all based on stochastic operations. Charniak et al. (2003) adapted this model with an entropy-based parser which enhanced the use made of syntactic information available to it. Gildea (2003) proposed a tree-to-tree alignment model in which subtree cloning was used to handle more reordering in parse trees.  
 590 Dependency treebanks have been used for Czech–English SMT by Čmejrek et al. (2003). Och et al. (2004) present and evaluate a wide variety of add-ons to a basic SMT system.

Another treebank-based approach to MT is the Data-Oriented Translation approach of Poutsma (2000) and Hearne and Way (2003). The authors consider this approach to be EBMT rather than SMT, and one could argue that with SMT taking on a more phrase-based and  
 595 syntax-based approach, while EBMT incorporates statistical measures of collocation and probability, the two approaches are quickly merging, a position argued by Way and Gough (2005).

## 6. Rapid development of MT for less-studied languages

An important attraction of corpus-based MT techniques is the possibility that they can be used  
 600 to quickly develop MT systems for less-studied languages (cf. [Article 23](#)), inasmuch as these MT techniques require only bilingual corpora and appropriate tools for alignment, extraction of linguistic data and so on. It must be said that some of the latest ideas, particularly in SMT, requiring treebanks and parsers make this less relevant. Nevertheless, empirical methods do seem to embody the best hope for resourcing under-resourced languages.

605 The first such attempt to demonstrate the feasibility of this was at the Johns Hopkins Summer Workshop in 1999, when students built a Chinese–English SMT system in one day (Al-Onizan et al. 1999). Although Chinese is not a less-studied language as such, it is of interest because English and Chinese are typologically quite dissimilar. The corpus used was the 7 million-word "Hong Kong Laws" corpus and the system was built using the EGYPT SMT  
 610 toolkit developed at the same workshop and now generally available online.

Germann (2001) tried similar techniques with rapidly developed resources, building a Tamil–English MT system by manually translating 24,000 words of Tamil into English in a six week period. Weerasinghe (2002) worked on Sinhala–English using a 50,000-word corpus from the World Socialist Web Site. Oard and Och (2003) built a system to translate between English  
 615 and the Philippine language Cebuano, based on 1.3m words of parallel text collected from five sources (including Bible translations and on-line and hard-copy newsletters). Foster et al. (2003) describe a number of difficulties in their attempt to build a Chinese–English MT system in this way.

## 7. Conclusions

620 MT is often described as the historically original task of Natural Language Processing, as well as the archetypical task in that it has a bit of everything, indeed in several languages; so it is no surprise that corpora – or at least collections of texts – have played a significant role in the

history of MT. However, it is only in the last 10–15 years that they have really come to the fore with the emergence and now predominance of corpus-based techniques for MT. This article has reviewed that history, from “reference corpora” in the days of rule-based MT via corpus-based translators’ tools, to MT methods based exclusively on corpus information. Many of the tools developed for corpus exploitation and described in other chapters in this book have had their genesis in MT, and research in corpus-based MT is certainly at the forefront of computational linguistics at the moment.

## 8. Acknowledgments

I would like to thank the editors and an anonymous reviewer for their very helpful comments on earlier drafts of this article. I would also like to thank Andy Way for his advice and suggestions on several sections of this article. All errors and infelicities remain of course my own responsibility.

## Literature

Al-Onizan, Y./Curin, J./Jahr, M./Knight, K./Lafferty, J./Melamed, D./Och, F.-J./Purdy, D./Smith, N.A./Yarowsky, D. (1999) Statistical machine translation: Final report, JHU Workshop 1999. Technical report, Johns Hopkins University, Baltimore, MD. Available at [http://www.clsp.jhu.edu/ws99/projects/mt/final\\_report/mt-final-report.ps](http://www.clsp.jhu.edu/ws99/projects/mt/final_report/mt-final-report.ps) [accessed 7 June 2005].

Alshawi, H./Srinivas, B./Douglas, S. (1998) Automatic acquisition of hierarchical transduction models for machine translation. In: *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, 41-47.

Arthern, P.J. (1978) Machine translation and computerized terminology systems: a translator’s viewpoint. In: Snell, B.M. (ed.) *Translating and the Computer: Proceedings of a Seminar, London, 14th November 1978*, Amsterdam (1979): North Holland, 77–108.

Barlow, M. (1995) ParaConc: A concordancer for parallel texts. In: *Computers and Texts* 10, 14–16.

Bowker, L. (2002) *Computer-Aided Translation Technology. A Practical Introduction*. Ottawa: University of Ottawa Press.

Brown, P. F./Cocke, J./Della Pietra, S. A./Della Pietra, V. J./Jelinek, F./Lafferty, J. D./Mercer, R. L./Roossin P. S. (1990) A statistical approach to machine translation. In: *Computational Linguistics* 16, 79–85; repr. in Nirenberg et al. 2003, 355–362.

Brown, P.F./Della Pietra, S. A./Della Pietra, V. J./Mercer, R. L. (1993) The mathematics of statistical machine translation: Parameter estimation. In: *Computational Linguistics* 19, 263–311.

Brown, R. D. (2000) Automated generalization of translation examples. In: *Proceedings of the 18th International Conference on Computational Linguistics, Coling 2000 in Europe*, Saarbrücken, Germany, 125-131.

- Brown, R. D. (2001) Transfer-rule induction for example-based translation. In: *MT Summit VIII Workshop on Example-Based Machine Translation*, Santiago de Compostela, Spain, 1-11.
- 665 Carl, M./Way, A. (2003) (eds) *Recent Advances in Example-Based Machine Translation*. Dordrecht: Kluwer Academic Press.
- Carl, M./Way, A. (2006/7) (eds) Special issue on example-based machine translation. *Machine Translation* 19(3-4) and 20(1).
- 670 Charniak, E./Knight, K./Yamada, K. (2003) Syntax-based language models for statistical machine translation. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA, 40-46.
- Church, K.W./Gale, W.A. (1991) Concordances for parallel texts. In: *Using Corpora, Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research*, Oxford, 40-62.
- 675 Cicekli, I. (2006) Inducing translation templates with type constraints. *Machine Translation* 19, 281-297.
- Cicekli, I./Güvenir, H.A. (1996) Learning translation rules from a bilingual corpus. In: *NeMLaP-2: Proceedings of the Second International Conference on New Methods in Language Processing*, Ankara, Turkey, 90-97.
- 680 Cicekli, I./Güvenir, H.A. (2003) Learning translation templates from bilingual translation examples. In: Carl & Way 2003, 255-286.
- Čmejrek, M./Cuřín, J./Havelka, J. (2003) Treebanks in machine translation. In: *Proceedings of The Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, Växjö, Sweden, 209-212.
- 685 Collins, B. (1998) *Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach*. PhD thesis, Trinity College, Dublin.
- Cranias, L./Papageorgiou, H./Piperidis, S. (1997) Example retrieval from a translation memory. In: *Natural Language Engineering* 3, 255-277.
- Dempster, A. P./Laird, N. M./Rubin, D. B. (1977) maximum likelihood from incomplete data via the EM algorithm. In: *Journal of the Royal Statistical Society Series B* 39, 1-38.
- 690 Foster, G./Gandraber, S./Langlais, P./Plamondon, P./Russel, G./Simard, M. (2003) Statistical machine translation: Rapid development with limited resources. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA, 110-117.
- 695 Foster, G./Langlais, P./Lapalme, G. (2002) User-friendly text prediction for translators. In: *2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, PA, 148-155.

- Garcia-Varea, I./Casacuberta, F./Ney, H. (1998) An iterative DP-based search algorithm for statistical machine translation. In: *Proceedings of the Fifth International Conference on Spoken Language Processing (ICSLP 98)*, Sydney, 1135–1139.
- 700 Gaussier, E./Langé, J.-M./Meunier, F. (1992) Towards bilingual terminology. In: *Proceedings of the ALLC/ACH Conference*, Oxford, 121–124.
- Germann, U. (2001) Building a statistical machine translation system from scratch: How much bang for the buck can we expect? In: *ACL-2001 Workshop on Data-Driven Methods in Machine Translation*, Toulouse, France, 1-8.
- 705 Germann, U./Jahr, M./Knight, K./Marcu, D./Yamada, K. (2001) Fast decoding and optimal decoding for machine translation. In: *Association for Computational Linguistics 39th Annual Meeting and 10th Conference of the European Chapter*, Toulouse, France, 228–235.
- Gildea, D. (2003) Loosely tree-based alignment for machine translation. In: *41st Annual Conference of the Association for Computational Linguistics*, Sapporo, Japan, 80-87.
- 710 Harris, B. (1988) Bi-text, a new concept in translation theory. In: *Language Monthly* 54, 8–10.
- Hearne, M./Way, A. (2003) Seeing the wood for the trees: data-oriented translation. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA, 165-172.
- 715 Isabelle, P. (1992a) Préface - Preface. In: *Quatrième Colloque international sur les aspects théoriques et méthodologiques de la traduction automatique, Fourth International Conference on Theoretical and Methodological Issues in Machine Translation, TMI-92*, Montréal, Canada, iii.
- 720 Isabelle, P. (1992b) Bi-textual aids for translators. In: *Screening Words: User Interfaces for Text, Proceedings of the 8th Annual Conference of the UW Centre for the New OED and Text Research*, Waterloo, Ont.; available at [http://rali.iro.umontreal.ca/Publications/urls/bi\\_textual\\_aids.ps](http://rali.iro.umontreal.ca/Publications/urls/bi_textual_aids.ps).
- 725 Jurafsky, D./Martin, J. H. (2000) *Speech and language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Upper Saddle River, NJ: Prentice Hall.
- Kay, M. (1980) The proper place of men and machines in language translation. Research Report CSL-80-11, Xerox PARC, Palo Alto, Calif.; repr. in *Machine Translation* 12 (1997), 3–23; and in Nirenberg et al. 2003, 221–232.
- 730 King, G. W. (1956) Stochastic methods of mechanical translation. *Mechanical Translation* 3(2), 38-39; repr. in Nirenburg et al. (2003), 37–38.
- Knight, K. (1999) Decoding complexity in word-replacement translation models. In: *Computational Linguistics* 25, 607–615.

- 735 Knight, K./Koehn, P. (2004) What's new in statistical machine translation? Tutorial at *HLT-NAACL 2004, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Alberta, Canada.
- Koehn, P./Knight, K. (2003) Feature-rich statistical translation of noun phrases. In: *41st Annual Conference of the Association for Computational Linguistics*, Sapporo, Japan, 311-318.
- 740 Koehn, P./Och, F. J./Marcu, D. (2003) Statistical phrase-based translation. In: *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Alberta, 127-133.
- Lehrberger, J. (1982), Automatic translation and the concept of sublanguage. In: Kittredge, R. I. & Lehrberger, J. (eds) *Sublanguage: Studies of Language in Restricted Semantic Domains*. Berlin: Mouton de Gruyter, 81-106; repr. in Nirenberg et al. 2003, 207-220.
- 745 Macdonald, K. (2001) Improving automatic alignment for translation memory creation. In: *Translating and the Computer 23: Proceedings from the Aslib Conference*, London [pages not numbered].
- Macklovitch, E./Russell, G. (2000) What's been forgotten in translation memory. In: White, J. S. (ed.) *Envisioning Machine Translation in the Information Future: 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000, Cuernavaca, Mexico*, Berlin: Springer, 137-146.
- 750 Manning, C. D./Schütze, H. (1999) *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Marcu, D./Wong, W. (2002) A phrase-based, joint probability model for statistical machine translation. In: *Conference on Empirical Methods for Natural Language Processing (EMNLP 2002)*, Philadelphia, PA.
- 755 McTait, K./Trujillo, A. (1999) A language-neutral sparse-data algorithm for extracting translation patterns. In: *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, Chester, England, 98-108.
- 760 Melby, A. (1981) A bilingual concordance system and its use in linguistic studies. In: Gutwinski, W. & Jolly, G. (eds) *LACUS 8: the 8th Lacus Forum, Glendon College, York University, Canada, August 1981*. Columbia, SC (1982): Hornbeam Press, 541-554.
- Nagao, M. (1984) A framework of a mechanical translation between Japanese and English by analogy principle. In Elithorn, A. & Banerji, R. (eds) *Artificial and Human Intelligence*, 765 Amsterdam: North-Holland, 173-180; repr. in Nirenberg et al. 2003, 351-354.
- Niessen, S./Vogel, S./Ney, H./Tillmann, C. (1998) ADP-based search algorithm for statistical machine translation. In: *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, 960-967.

- 770 Nirenburg, S./Domashnev, C./Grannes, D.J. (1993) Two approaches to matching in example-based machine translation. In: *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation TMI '93: MT in the Next Generation*, Kyoto, Japan, 47–57.
- 775 Nirenberg, S./Somers, H./Wilks, Y. (2003) (eds) *Readings in Machine Translation*. Cambridge, Mass.: MIT Press.
- Oard, D. W./Och, F. J. (2003) Rapid-response machine translation for unexpected languages. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*, New Orleans, USA, 277-283.
- 780 Och, F.J./Gildea, D./Khudanpur, S./Sarkar, A./Yamada, K./Fraser, A./Kumar, S./Shen, L./Smith, D./Eng, K./Jain, V./Jin, Z./Radev, D. (2004) A smorgasbord of features for statistical machine translation. In: *Human Language Technology Conference and Annual Meeting of the North American Chapter of the Association for Computational Linguistics*, Boston, MA, 161-168.
- 785 Och, F. J./Ney, H. (2003) A systematic comparison of various statistical alignment models. In: *Computational Linguistics* 29, 19–51.
- Och, F. J./Ney, H. (2004) The alignment template approach to statistical machine translation. In: *Computational Linguistics* 30, 417–449.
- 790 Och, F. J./Tillmann, C./Ney, H. (1999) Improved alignment models for statistical machine translation. In: *Proceedings of the 1999 Joint SIGDAT Conference of Empirical Methods in Natural Language Processing and Very Large Corpora*, College Park, MD, 20–28.
- Och, F. J./Ueffing, N./Ney, H. (2001) An efficient A\* search algorithm for statistical machine translation. In: *Proceedings of the Data-Driven Machine Translation Workshop, 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, 55–62.
- 795 Planas, E./Furuse, O. (1999) Formalizing translation memories. In: *Machine Translation Summit VII*, Singapore, 331-330; repr. in Carl & Way 2003, 157-188.
- Poutsma, A. (2000) Data-oriented parsing. In: *COLING 2000 in Europe: The 18th International Conference on Computational Linguistics*, Luxembourg, 635-641.
- 800 Rapp, R. (2002) A part-of-speech-based search algorithm for translation memories. In: *LREC 2002, Third International Conference on language Resources and Evaluation*, Las Palmas de Gran Canaria, Spain, 466-472.
- Romary, L./Mehl, N./Woolfs, D. (1995) The Lingua parallel concordancing project: Managing multilingual texts for educational purposes. In: *Text Technology* 5, 206–220.
- 805 Sato, S./Nagao, M. (1990) Toward memory-based translation. In: *COLING-90, Papers Presented to the 13th International Conference on Computational Linguistics*, Helsinki, Finland, 3, 247–252.

- Simard, M./Foster, G./Perrault, F. (1993) TransSearch: A bilingual concordance tool. Industry Canada Centre for Information Technology Innovation (CITI), Laval, Canada, October 1993; available at <http://rali.iro.umontreal.ca/Publications/urls/sfpTS93e.ps>.
- 810 Somers, H. (2003) Translation memory systems. In: Somers, H. (ed.) *Computers and Translation: A Translator's Guide*, Amsterdam: Benjamins, 31-47.
- Somers, H./Fernández Díaz, G. (2004) Translation memory vs. example-based MT: What is the difference? In: *International Journal of Translation* 16(2), 5–33; based on: Diferencias e interconexiones existentes entre los sistemas de memorias de traducción y la EBMT. In: 815 Corpas Pastor, G. & Varela Salinas, M.a-J. (eds) *Entornos informáticos de la traducción profesional: las memorias de traducción*, Granada (2003): Editorial Atrio, pp. 167–192.
- Somers, H./Tsuji, J./Jones, D. (1990) Machine translation without a source text. In: *COLING-90, Papers Presented to the 13th International Conference on Computational Linguistics*, Helsinki, Finland, 3, 271-276; repr. in Nirenberg et al. 2003, 401–406.
- 820 Sumita, E./Iida, H./Kohyama, H. (1990) Translating with examples: A new approach to machine translation. In: *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*, Austin, Texas, 203–212.
- Tillmann, C./Ney, H. (2003) Word reordering and a dynamic programming beam search 825 algorithm for statistical machine translation. In: *Computational Linguistics* 29, 97–133.
- Tillmann, C./Vogel S./Ney, H./Sawaf, H. (2000) Statistical translation of text and speech: First results with the RWTH system. In: *Machine Translation* 15, 43–74.
- Tillmann, C./Vogel S./Ney, H./Zubiaga, A. (1997) A DP-based search using monotone 830 alignments in statistical translation. In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, 289–296.
- Ueffing, N./Och, F. J./Ney, H. (2002) Generation of word graphs in statistical machine translation. In: *Conference on Empirical Methods for Natural Language Processing (EMNLP 2002)*, Philadelphia, PA, 156–163.
- 835 Wang, Y.-Y./Waibel, A. (1997) Decoding algorithm in statistical machine translation. In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, Spain, 366–372.
- Wang, Y.-Y./Waibel, A. (1998) Modeling with structures in statistical machine translation. In: 840 *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Canada, 1357-1363.
- Watanabe, T./Sumita, E. (2003) Example-based decoding for statistical machine translation. In: *MT Summit IX, Proceedings of the Ninth Machine Translation Summit*, New Orleans, 845 USA, 410-417.

Way, A./Gough, N. (2005) Comparing example-based and statistical machine translation. In: *Journal of Natural Language Engineering* 11, 295-309.

850 Weerasinghe, R. (2002) Bootstrapping the lexicon building process for machine translation between 'new' languages. In: Richardson, S. D. (ed.) *Machine Translation: From Research to Real Users, 5th Conference of the Association for Machine Translation in the Americas, AMTA2002, Tiburon, CA*, Berlin: Springer, 177-186.

Wu, D. (1997) Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. In: *Computational Linguistics* 23, 377-403.

855 Wu, D./Wong, H. (1998) Machine translation with a stochastic grammatical channel. In: *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Canada, 1408-1414.

860 Yamada, K./Knight, K. (2001) A syntax-based statistical translation model. In: *Association for Computational Linguistics 39th Annual Meeting and 10th Conference of the European Chapter*, Toulouse, France, 523-530.