

# Bilingual Parallel Corpora and Language Engineering

Harold Somers

*Department of Language Engineering,*

*UMIST, Manchester, England*

harold@ccl.umist.ac.uk

## 1. Introduction

The use of corpora has become an important issue in Language Engineering (LE). In this paper we will be considering a specific type of corpus, the **bilingual parallel corpus**. By “parallel corpus”, we mean a text which is available in two (or more) languages: it may be an original text and its translation, or it may be a text which has been written by a consortium of authors in a variety of languages, and then published in various language versions. A corpus of this type of text is sometimes called a “comparable corpus”, though this term is also used (confusingly) for a corpus of similar but not necessarily equivalent texts. Another term sometimes found is “bitext”, due to Brian Harris (1988).

Parallel corpora are a valuable source of a kind of linguistic metaknowledge, which forms the basis of techniques such as tokenization, POS-tagging, morphological and syntactic analysis, which in turn can be used to develop LE applications.

This paper focuses on problems (and solutions) related to the extraction of linguistic meta-knowledge from parallel corpora.

## 2. “First, catch your corpus”

The first requirement for knowledge extraction from bilingual corpora is, rather obviously, a parallel corpus. Fully annotated aligned multilingual parallel corpora in a number of languages are becoming increasingly widely available through various coordinated international efforts. A visit to any of a number of websites devoted to corpora in general and bilingual corpora in particular reveals a long list of such collections. The W3C website at Essex University ([clwww.essex.ac.uk/w3c/general.html](http://clwww.essex.ac.uk/w3c/general.html)) is a good starting point.

Nevertheless, even though the number of collections is ever increasing, the number of different languages featured is still rather small. Also, some of the collections are relatively unfocused in terms of subject matter. In either case there may be a problem of coverage for a particular need. In this case, you might need to attempt to locate and analyse your corpus from scratch. So we begin by considering some ways of automatically locating parallel texts, and some issues involved in retrieving and storing such data.

### 2.1. Locating parallel corpora automatically

Although English is overwhelmingly the *lingua franca* of the World Wide Web, a great number of web sites have parallel material in several languages. These evidently provide an instant source of parallel texts, if they can be located and successfully aligned.

Interesting work on automatically identifying and locating parallel corpora has been initiated by Resnik (1998, 1999).

The idea is first of all to find likely candidate pairs of texts using such “tricks” as searching for sites which seem to have parallel “anchors” (see below), often accompanied by images of flags, or pairs of filenames which differ only in the identification of a language, e.g. with alternative directories in the paths, or suffixes such as *.en* and *.fr*.

These candidates are then evaluated by comparing, in a very simplistic manner, their content: since they are usually HTML documents, it is usually quite easy to align the HTML mark-up (heading and subheading identifiers, for example), and to compare the amount of text between each anchor. In this way, we get a rough map of the structures of the two documents. These can then be compared using a variety of more or less sophisticated techniques which may or may not include the kinds of linguistic methods used in the alignment of known parallel texts – see next section. Flexibility in mark-up conventions can undermine this technique, however. For example, Figure 1 shows parallel English and French pages (written by the current author) with minor differences in mark-up and content.

<pre> &lt;HTML&gt;   &lt;HEAD&gt;&lt;TITLE&gt;ATLAS Symposium&lt;/TITLE&gt;&lt;/HEAD&gt;   &lt;BODY bgcolor="ffffff" text="115511" LINK="004080" vLINK="0040800"&gt;     &lt;center&gt;&lt;img src="..." alt=logo height=145 width=184 &gt;     &lt;h1&gt;Arabic Translation and Localisation Symposium&lt;p&gt;Symposium sur la Traduction et la Localisation en Arabe&lt;br&gt;&lt;img src="arabatlas.gif" alt=arabic&gt;&lt;/h1&gt;     ...   &lt;/center&gt;   &lt;p&gt;It is one of the five official languages of the United Nations, it has 260 million native speakers, and is used as a second language by a further 1.3 billion people.   ...   &lt;center&gt;   &lt;li&gt;Arabic corpus processing   &lt;li&gt;Development of Arabic resources   &lt;li&gt;Web tools for Arabic </pre>	<pre> &lt;HTML&gt;   &lt;HEAD&gt;&lt;TITLE&gt;Symposium ATLAS&lt;/TITLE&gt;&lt;/HEAD&gt;   &lt;BODY TEXT="#115511" BGCOLOR="#FFFFFF" LINK="#004080"VLINK="#048000"&gt;     &lt;CENTER&gt;&lt;IMG SRC="..." ALT="logo" HEIGHT=145 WIDTH=184&gt;     &lt;H1&gt;Symposium sur la Traduction et la Localisation en Arabe&lt;P&gt;Arabic Translation and Localisation Symposium&lt;BR&gt;&lt;IMG SRC="arabatlas.gif"&gt;&lt;/h1&gt;     ...   &lt;/CENTER&gt;   &lt;p&gt;L'une des cinq langues officielles de l'ONU est l'&lt;B&gt;arabe&lt;/B&gt;, la languematernelle de 260 millions de locuteurs, qu'utilisent environ 1.3 milliards de musulmans comme deuxi&amp;egrave;me langue.   ...   &lt;CENTER&gt;   &lt;LI&gt;les standards de codage des caract&amp;egrave;res arabes&lt;/LI&gt;   &lt;LI&gt;le traitement des corpus en arabe&lt;/LI&gt;   &lt;LI&gt;le d&amp;eacute;veloppement des ressources pour l'arabe&lt;/LI&gt;   &lt;LI&gt;les outils Internet pour l'arabe&lt;/LI&gt; </pre>
---	---

Figure 1. HTML versions of parallel web pages. Notice differences in capitalization in the tags, order of elements in the <BODY> tag, and textual differences, e.g. an additional <LI> item in the French version.

## 2.2.Storage and encoding

Having located a suitable parallel corpus, there remain a number of aspects to consider before the process of linguistic knowledge extraction can begin. One, which should not be ignored is the issue of determining the legal position with respect to the text: even though the WWW is a kind of public domain, much of the text found on it is nevertheless subject to **copyright** and ownership. It is thus important to obtain the owners' agreement before using the material, especially if, as we will suggest below, you intend to make the data available through various data sharing initiatives. Some suggestions on these legal issues are offered in Thompson (2000).

Another issue – also addressed by Thompson (2000) – is obvious if you are interested in a language pair which use a different script, namely the question of **character sets**. Although for some languages there are uniquely recognised standards for web pages, for others there are conflicting standards, or no standards at all. For some languages, web pages turn out to be graphic images.

Another issue that has been widely addressed is the question of **encoding**. This term refers to annotations which are added to the text in order to facilitate data manipulation and analysis. The type of “mark-up” that can be added to the text ranges from identification information such as its source, subject matter, language, date of capture, and so on, to more linguistic mark-up as the result of analysis: part-of-speech (POS) tags, document structure codes and so on. There have been various standardisation efforts in this area, most notably the Text Encoding Initiative (TEI), whose guidelines and recommendations (Sperberg-McQueen & Burnard, 1994) seem to have been widely accepted. Standardisation facilitates the exchange of data, and equally, there have been numerous movements towards sharing and exchanging corpus material.

## 3. Alignment

In order to extract information from our parallel text, it is first necessary to **align** the two texts at some global level, typically paragraph or sentence. By “align” is meant the association of chunks of text in the one document with their translation or equivalent text in the other document. Some approaches to alignment involve the use of some sort of traditional analysis of the texts (e.g. parsing, tagging, or the use of bilingual lexicons), while others take an entirely automatic approach. For our purposes – i.e. extraction of linguistic knowledge – the former would seem to involve a kind of vicious circle, since they make use of precisely the kind of information we are trying to extract. There is a huge and continually growing body of literature on this subject. The paragraphs that follow are not intended as an exhaustive review – see Wu (2000) for a more detailed review.

Gale & Church (1991a) and Brown et al. (1991) both developed alignment programs based on the simple assumption that there is a significant correlation in the **relative length** of texts which are translations of each other. The former measured length in characters, the latter in words. Simard et al. (1992) suggested that such alignments could be improved upon if apparent cognates in the two texts were used as aids (see below). Gale & Church (1991b) took the output of their alignment program and used it to identify correspondences at the word level. Much of the early success of all these approaches was no doubt due to the fact that the universally used Canadian Hansard corpus was very “well-behaved” in that most of the sentences and paragraphs lined up

nicely, and also syntactically and lexically French and English are quite similar. McEnery & Oakes (1996) illustrate widely varying success rates for different language pairs and domains. Figure 2 illustrates the problem in a simple way. Consider two texts with segments of length  $\langle 10,30,15,12 \rangle$  and  $\langle 12,14,15,12,11,2 \rangle$ . At first sight, the most intuitive alignment pairs up the segments of nearly equal length, as indicated by the arrows. But a more likely alignment, which accounts for more of the text, is shown by the balloons, where the segment of length 30 is aligned in 1:2 mapping with the 14+15 segments, and so on.

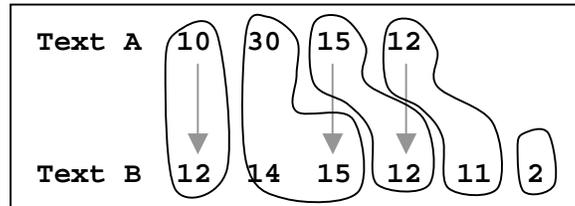


Figure 2. Alignment by segment length.

Inspired by Simard et al. (1992), Church (1993) addressed the problem of “noisy” texts by trying to align on the basis of similar short stretches of characters (**cognates**). Other researchers using cognates include Johansson et al. (1993) and Melamed (1996a). For languages using different writing systems, this technique applies in a much more limited way, inasmuch as texts share identifiable character strings such as numbers, or the mapping between the writing systems is straightforward.

Church looked at texts which had been scanned in from hard copy, and so contained misalignment problems caused by different pagination, e.g. a footnote suddenly appearing in the middle of a paragraph, or figure headings out of sequence. Similarly, Chen (1993) overcame the noise problem by aligning sentences on the basis of word alignments.

Most of the approaches have in common a technique which involves identification of **anchor points** and verification of the comparability of the textual material between the anchors. These anchors can, in the simplest case, be structural, as in early work by Gale & Church (1991a) where sentence boundaries are taken to make an initial segmentation. Then, certain types of alignment across sentence boundaries are permitted and quantified (e.g. where two sentences in one text are merged in the translation, or vice versa), with all possible alignments being compared using dynamic programming techniques.

Alternatively, and quite commonly, translation **word-pairs** are taken as the anchor points. This alignment at the word level is often an end-goal in itself, as exemplified by the pioneering work of Kay and Röscheisen (1993). Their algorithm performs the two functions of sentence alignment and word alignment simultaneously, the one feeding off and reinforcing the other.

A quite widely used idea is illustrated in Figure 3. We first identify anchor-points throughout the text (a), and then pick those that are closest to the ideal alignment which is a diagonal line (b). These then define subtext regions (c) in which the process can be iterated. Various “smoothing” techniques can be used to reduce the search space even further (d).

Apart from automatic estimation of translation pairs, a number of sentence alignment algorithms rely on **machine-readable dictionaries** as a method for finding

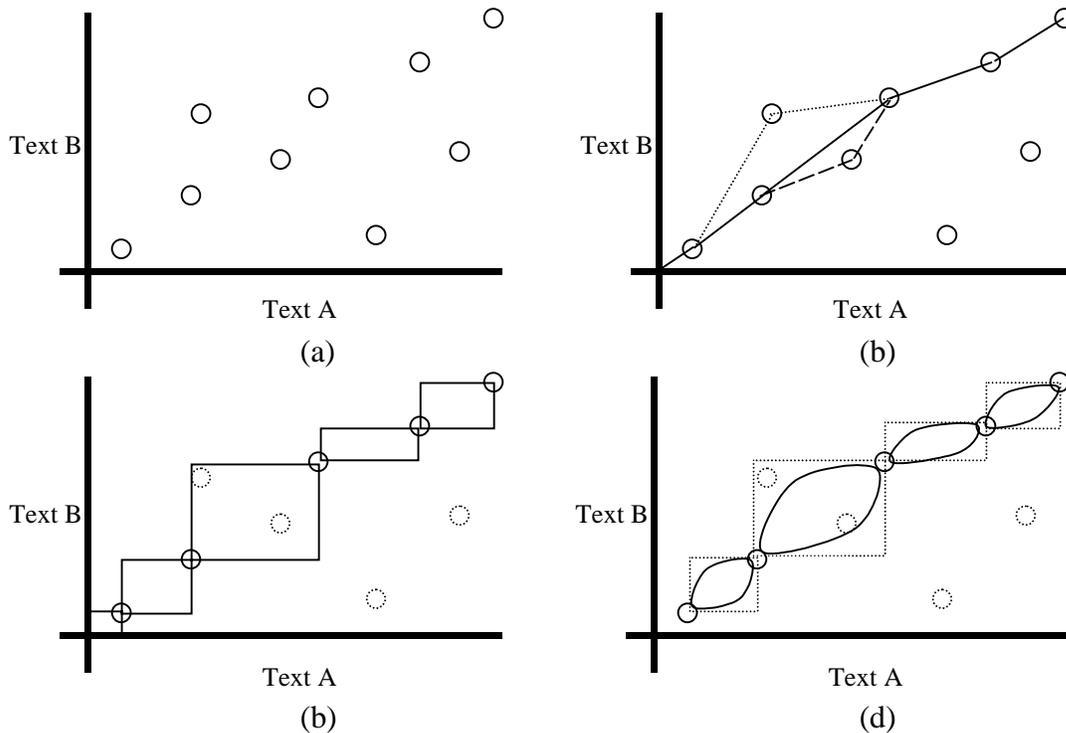


Figure 3. Stages in anchor-based alignment.

lexical anchor points. This technique of course relies on the availability of a suitable dictionary, not to mention the need for efficient lemmatization in the case of highly inflected languages. Again, this seems like a vicious circle if the aim of the alignment is to extract vocabulary; but as we will see below, aligned parallel corpora can be used for the extraction not of everyday vocabulary, but of domain-specific lexical pairings, notably novel terminology and, especially where different writing systems are involved, transliterations of proper names.

## 4. Extraction of bilingual vocabulary and terminology

Algorithms for extracting bilingual vocabulary from aligned parallel corpora exploit – or depend on – the following characteristics of translated texts (Fung & Yee, 1998:2):

- Words have one sense per corpus.
- Words have a single translation per corpus.
- There are no missing translations in the target document.
- The frequencies of words and their translations are comparable.
- The positions of words and their translations are comparable.

The algorithms that we will be discussing take advantage of the above facts, and their principle weaknesses are the extent to which the above do *not* hold. We will return to this point later.

### 4.1. Identifying likely word-pairs

One of the earliest attempts to extract bilingual vocabulary from a parallel corpus was Kay and Röscheisen (1993). As mentioned above, their method was a hybrid of sentence-

and word-alignment. Their technique is to find word pairs which are most probably alignable on the basis of similar **distribution**. This distribution is defined in terms of text sectors, and Dice's coefficient is used to quantify the probability. Dice's coefficient (1) is a simple calculation which compares  $c$ , the number of times the two candidate words occur in the same sector with  $a$  and  $b$ , the number of times the source or target words occur independently.

$$(1) \quad \text{Dice} = \frac{2c}{a+b}$$

The algorithm is iterative in that the sentences containing high-scoring word pairs are established as anchors which allow the text to be split up into smaller segments, affording more and more results.

Another distribution-based algorithm is K-vec (Fung & Church, 1994). In this case, the parallel texts are split into  $K$  equal-sized segments and the distributions of each word are recorded in binary vectors  $1 \dots K$ . The binary vectors for two candidate words  $w_s$  and  $w_t$ , are then compared, the similarity of any two distributions being quantified using two measures, Mutual Information (MI) and a t-score, as in (2)–(3).

$$(2) \quad \text{MI} = \log_2 \frac{aK}{bc}$$

$$(3) \quad t = \frac{b}{K \sqrt{\frac{a}{K^2}}}$$

Fung & Church (1994) initially reported good results with this algorithm for alignment of noisy French–English texts, but Jones & Somers (1995) reported less impressive results with this algorithm with English, Japanese and German texts. Fung & McKeown (1997a) also reported the poor performance of the K-vec algorithm with Japanese–English and Chinese–English parallel corpora, and proposed a weighted MI, calculated as in (4).

$$(4) \quad \text{wMI} = \frac{a}{K} \log_2 \frac{aK}{(a+b)(a+c)}$$

Gao (1997) noted that in addition to the Dice coefficient, other similarity measures widely used in Information Retrieval (IR) could also be applied in this case, including the Jaccard coefficient (5) and the Cosine coefficient (6). Gao tested the various measures with his parallel English–Chinese text, and found the Jaccard the best measure.

$$(5) \quad \text{Jaccard} = \frac{c}{a+b-c}$$

$$(6) \quad \text{Cosine} = \sqrt{\frac{c}{ab}}$$

Recognising some flaws in the performance of K-vec, Fung & McKeown (1997a) developed DK-vec: like K-vec, it tries to recognise translation pairs by considering their distribution. This time, the distribution is expressed as **recency vectors**, i.e. sequences of integers representing the gap, in characters, between each instance of a word. Figure 4 shows the recency vectors for the Chinese word for ‘governor’ compared to *governor*, *president* and *bill* in the Hong Kong Hansard corpus. The similarity of the “trace” for the

two words which are related is very graphic. The vectors are then compared using the Dynamic Time Warping algorithm, although simpler distance measures such as Minimal Edit Distance or Levenshtein distance are used by Dagan (1996) and Somers (1998) in replications of Fung & McKeown’s work with different language pairs.

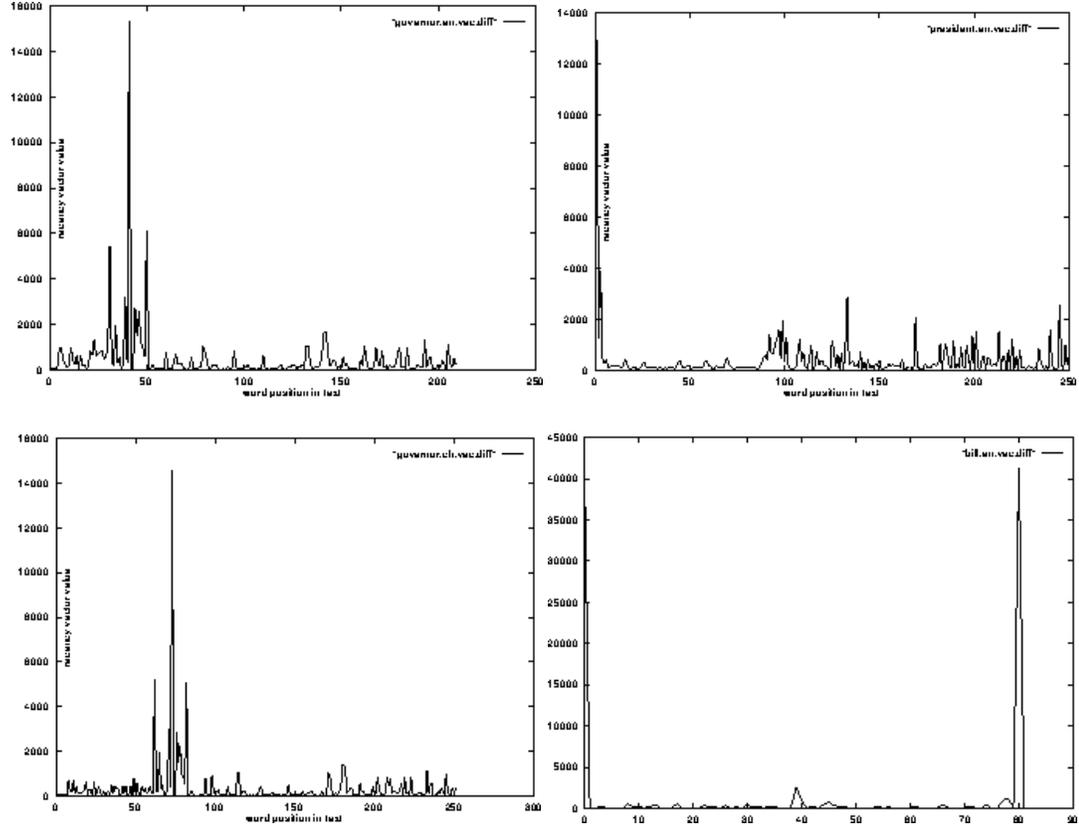


Figure 4. Recency vectors for the Chinese word for ‘governor’ and the English words governor, president and bill in the Hong Kong Hansard corpus, from Fung & McKeown (1997a).

#### 4.2. Extracting terminology: identifying multiword equivalents

One major drawback to all the techniques described in the previous section is the necessary assumption that word equivalents are on a 1:1 basis. Apart from the fact that this is generally not always true in languages (even making the prior assumption that we know what a word is!), it is especially unhelpful if our aim is to identify bilingual terminology. It is in this endeavour that parallel corpus work comes to the fore: often, the goal of extracting parallel vocabulary is undermined somewhat by the existence of machine-readable bilingual dictionaries. Specialist terminology, however, is almost always absent from such resources, and bilingual parallel corpora become the primary – perhaps only – source of such material.

Searching for multiword terms in a parallel corpus introduces a further aspect of word distribution which can be addressed by statistical means: considering the corpora independently, we can search for likely terms by looking for **collocations**, i.e. sequences of words which co-occur frequently and – if we are lucky – tend not to occur on their own.

There is a considerable literature on collocations, e.g. Church & Hanks, 1989; Smadja, 1993), though we should perhaps distinguish “loose collocations” – sets of words which typically co-occur in a text and can be used, for example, for indexing or retrieval purposes in IR (e.g. a text containing words like *doctor* and *nurse* may be about hospitals), or to disambiguate polysemous words (e.g. if we want to know which meaning of *bar* is intended, we can look for words like *drink*, *beer*, *bartender*, or *lawyer*, *judge*, *chambers*) – and “contiguous collocations” which might be candidates for terminology. Gaussier et al. (2000) discuss how collocation information can also be used to identify possible technical terms. Often with minority languages terminology actually needs to be *created*. To take an example from the UK situation, Urdu does not (or did not) have words for ‘employment benefit’ or ‘poll tax’, both British concepts which needed to be translated for some community information leaflet. The danger is that translators will simply invent a word or just transliterate it into the Urdu writing system. If terminology is created locally by translators in this manner, there is the risk of a proliferation of competing terms which of course can lead to confusion.

There is a range of measures for (monolingual) collocations. The  $z$ -score is perhaps the most familiar: it quantifies the collocational force of one word  $w_i$  with respect to another  $w_j$  as in (7),

$$(7) \quad z = \frac{O - E}{\sigma}$$

where  $O$  is the observed frequency of  $w_i$  co-occurring with  $w_j$  (in close proximity, or contiguous with it, as the case may be),  $E$  the expected frequency of  $w_i$ , and  $\sigma$  is the standard deviation of occurrence of  $w_i$  in the whole text as given by (8),

$$(8) \quad \sigma = \sqrt{N(p(1-p))}$$

where  $p$  is the probability of occurrence of  $w_i$ , and  $N$  is the total number of word tokens in the text.

Other measures used are MI and t-score as already seen above (2), (3). Further alternatives have been proposed. Gale & Church (1991c) introduced the  $\Phi^2$  coefficient (9), Dunning (1993) proposed the loglike coefficient  $G^2$  (10), while Daille (1995a,b) used two measures closely related to MI, the cubic association ratio (11) and the association ratio IM (12). Kitamura & Matsumoto (1996) proposed (13) a modification of Dice’s coefficient weighted to take into account the cooccurrence frequency.

$$(9) \quad \Phi^2 = \frac{(ad - bc)^2}{(a+b)(a+c)(b+c)(b+d)}$$

$$(10) \quad G^2 = f(a) + f(b) + f(c) + f(d) - f(a+b) - f(a+c) - f(b+d) - f(c+d) + f(a+b+c+d)$$

where  $f(x) = x \log x$

$$(11) \quad \text{MI}^3 = \log_2 \frac{a^3}{(a+b)(a+c)}$$

$$(12) \quad \text{IM} = \log_2 \frac{a}{(a+b)(a+c)}$$

$$(13) \quad \text{sim}(w_i, w_j) = \log_2 c \frac{2c}{(a+b)}$$

All of these measures have drawbacks however, and there are a number of studies which compare them and try to mitigate their weaknesses. It should also be noted that not all collocations are terms. A number of studies take advantage of what is known about the structure of terms to help try to extract terminology automatically. For example, Daille et al. (1994) and Daille (1996) look for appropriate sequences of POS tags as well as recurring word sequences, recognising that terms (in French) are often of the form *N de N*. Justeson & Katz (1995) similarly incorporate knowledge about typical term structure into their algorithm.

Once candidate terms have been determined monolingually, we can turn our attention to identifying their translation equivalents in the parallel corpus. Many of the experiments reported seem to take the same basic approach, namely identifying possible terms monolingually and then searching for their translation in an aligned parallel corpus.

Gaussier & Langé (1997) describe two methods for assessing the translations of terms found monolingually. The first method compares a number of measures like MI,  $\Phi^2$  and so on, that we have already seen. The second method relies more on the previously calculated scores for single-word alignment. Following up on that work, McEnerey et al. (1997) incorporated the use of cognates (see above) to improve their results.

Smadja et al. (1996) describe a program named *Champollion* which takes a parallel corpus which has been sentence-aligned, uses Xtract (Smadja, 1993) to identify possible collocations, and then uses a method very similar to K-vec to evaluate the probability that collocations thus identified are translations of each other. They used Dice's coefficient and MI to quantify the matches, and experimented with data from the Canadian Hansard corpus.

Dagan and Church (1997) proposed a semi-automatic tool, *termight* for constructing bilingual glossaries. Like *Champollion*, the task is divided into two parts: identifying monolingual term lists, and then finding their translations. The first step is closely modelled on the approach of Justeson & Katz (1995), combining frequent cooccurrence with appropriate syntactic pattern as a criterion. Importantly, the human terminologist has a major role in the process, so much of the effort in *termight* is focussed on presenting the results in a user-friendly manner. The second step relies on the fact that the bilingual corpora have been aligned both at sentence and at word level. Even though this alignment might be flawed, it means that for every word sequence identified as a possible term monolingually, there is a corresponding sequence of aligned words in the parallel corpus. Again the human terminologist has a role to play here. One advantage of this method is that it deals well with the typical word-order scrambling effect found especially with language pairs like English and French (e.g. *optional parameters box* corresponds to *zone paramètres optionnels*), which Dagan & Church use to test their program.

### 4.3. How good are these algorithms?

The success rates of the algorithms described here varies tremendously. Somers (1998) has identified some of the factors impinging on one particular algorithm, DK-vec, including obvious factors such as genre, language-pair, amount and homogeneity of data. As mentioned above, most of the algorithms make certain assumptions about the nature of parallel corpora, and it is useful to revisit those assumptions here to see what their effect is.

*Words have one sense per corpus*

This is the basic assumption underlying the “sublanguage” approach to natural language processing. It is often true, especially for words which have terminological status; but homonymy is not avoidable, even in narrow domains.

*Words have a single translation per corpus*

This is a much less safe assumption, which is particularly undermined by the fact that inflectional morphology and compounding methods differ from language to language. The assumption of 1:1 word correspondence is of course naive, bearing in mind polysemy and homonymy, and the converse problem of translation divergence (e.g. German has two competing terms for ‘computer’, *Rechner* and *Computer*). The assumption is undermined further by the fact that local syntactic conditions might result in inflectional morphology in one language but not the other: in particular, the distribution of singular and plural can differ widely between otherwise closely related languages, without even considering grammatical case and gender. Where possible, this can be overcome by subjecting the corpora to a process of **lemmatization**.

Another problem is that multi-word compounds in one language may correspond to what are typographically single words in another. This problem has been discussed for German (Jones & Alexa, 1997), and for Swedish (Ahrenberg et al., 1998). And for languages such as Chinese, Japanese and Korean, the problem is further compounded by the fact that their writing systems do not mark word-boundaries explicitly, so a prior step in any word-alignment task is always word segmentation, which may introduce a certain amount of error, cf. Gao (1997), Wu & Xia (1994), Shin et al. (1996).

*There are no missing translations in the target document*

This is a somewhat safer assumption, but not entirely so. It is not unusual to find that some portion of the source text has been omitted in the target text, either through carelessness, or because it does not apply to the target-language readership. Interestingly, one off-shoot of work on alignment has been the development of tools to help translators check for missing text in translations (Melamed, 1996b).

*The frequencies of words and their translations are comparable*

The main problem with this assumption is again the fact that a single word in one language can have a variety of translations in the other just because of grammatical inflection. Somers (1998:130) gives the example of the word *all* occurring 40 times in a certain English corpus, while in the corresponding Spanish corpus we get *todas* 25 times, *todos* 19, *todo* 5, and *toda* once. Another source of discrepancy is the use of capitalisation, especially comparing, say, English with German (where all nouns are capitalised irrespective of their position in the sentence), or Japanese (which does not distinguish upper and lower case).

*The positions of words and their translations are comparable*

This seems to be the most fundamental of assumptions in alignment. The extent to which it is true depends on the granularity of the alignment. Clearly, at sentence level it is hardly true at all: word-order is a fundamental difference between many languages, not just the obvious case of, say, adjectives preceding or following the noun, but also the relative order of main and subordinate clauses (*A because B* vs. *B and so A*, for example). But as the size of the text element being considered grows, this effect becomes minimised. For some language pairs, there remains a certain amount of “scrambling”. For example, Gao (1997) reports that minor omissions and changes in the order of

presentation of material was a major feature of his English–Chinese parallel corpus, taken from a Taiwanese current affairs magazine.

We might end this section by noting the pioneering work by Fung and associates (Fung, 1995, 1998; Fung & McKeown, 1997b; Fung & Yee, 1998) on bilingual lexicon extraction from *non*-parallel corpora. Here the attention turns to corpora which are not translations of each other, but are “comparable” corpora, i.e. collections of texts of the same genre, covering the same domain, and so on. Clearly, the assumptions we have just discussed are not applicable to non-parallel comparable corpora. Individual word distributions and frequencies, and the possibilities of lexical alignment are obviously not available in this scenario. But the corpus linguist *can* look for other useful patterns, notable comparable contexts and usage. Fung’s algorithms owe a lot to IR techniques, in particular measures of distribution (term frequency), and similarity of context (IDF). This seems to be a promising new departure, and one where the techniques of corpus linguistics and IR can come together.

## 5. Extracting translation templates

Besides bilingual vocabulary and terminology, aligned parallel corpora have been used to extract translation templates, especially for the purposes of Example-based Machine Translation (EBMT) and the related translator’s tool, Translation Memory.

EBMT was first proposed as long ago as 1981 (Nagao, 1984), but was only developed from about 1990 onwards – see Somers (1999) for an extensive overview. The basic idea is to reuse examples of already existing translations as the basis or model for a new translation. In its basic form, the examples in EBMT are stored as pairs of aligned text fragments, usually sentences, though in some implementations stored examples are annotated with POS tags or other information, or even stored as linked tree structures. Translation proceeds by first matching the input with a suitable example, then adapting the example to the new case.

A quite popular technique is to try to **generalize** translation templates from stored examples. This can be illustrated by considering example (14a), from Brown (1999), which can be generalized as (14b) or (14c), thus facilitating the identification of its translation as a model for the translation of (15).

- (14) a. John Miller flew to Frankfurt on December 3rd.  
 b. <1stname> <lastname> flew to <city> on <month> <ord>.  
 c. <person-m> flew to <city> on <date> .

- (15) Dr Howard Johnson flew to Ithaca on 7 April 1997.

Furuse & Iida (1992) were perhaps the first to propose the idea, though in their case the generalization was done manually. Collins & Cunningham (1995) showed how examples could be generalized for the purposes of retrieval, though with a corresponding trade-off in precision and recall. Kaji et al. note that the process can become quite complicated. For example, the English–Japanese cases in (16a,b) might lead you to propose a generalized template (16c). But the counterexamples in (17) show that the templates must be refined, as in (18) to give more information.

- (16) a. play baseball → *yakyu o suru*  
 b. play tennis → *tenisu o suru*

- c. play X[NP] → X[NP] o suru  
 (17) a. play the piano → *piano o hiku*  
 b. play the violin → *baiorin o hiku*  
 c. play X[NP] → X[NP] o hiku  
 (18) a. play X[NP/sport] → X[NP] o suru  
 b. play X[NP/instrument] → X[NP] o hiku

Carl (1999) similarly refines examples to give generalizations based on syntactic categories and morphological features. Likewise, Langé et al. (1997) describe their “skeleton-sentences” approach to Translation Memories, where candidates for generalization are term pairs or “transwords” (roughly, alphanumeric and proper names which are not translated). Other researchers reporting similar ideas include Nomiya (1992), Almuallim et al. (1994), Akiba et al. (1995), Jain et al. (1995), and Matsumoto & Kitamura (1995).

Several researchers have tried to extract such templates automatically, notably for English–Turkish (Cicekli & Güvenir, 1996; Güvenir & Tunç, 1996; Güvenir & Cicekli, 1998) and for English–Spanish (McTait et al., 1999; McTait & Trujillo, 1999). For example, from the English–Turkish pairs in (19), the lexical pairings *ticket* ⇔ *bilet* and *pen* ⇔ *kalem* can easily be inferred, along with the template in (20). Similarly, the examples in (21) permit the extraction of the templates in (22).

- (19) a. I took a ticket from Mary ⇒ *Mary'den bir bilet aldım.*  
 b. I took a pen from Mary ⇒ *Mary'den bir kalem aldım.*  
 (20) I took a ... from Mary ⇒ *Mary'den bir ... aldım.*  
 (21) a. The Commission gave the plan up ⇒ *La Comisión abandonó el plan.*  
 b. Our Government gave all laws up ⇒ *Nuestro Gobierno abandonó todas las leyes.*  
 (22) a. ... gave ... up ⇒ *abandonó*  
 b. The Commission ... the plan ⇒ *La Comisión ... el plan*  
 c. Our Government ... all laws ⇒ *Nuestro Gobierno ... todas las leyes*

More examples might permit you to break (19) and (22b,c) down further and identify which parts can be coupled (you may or may not make the assumption that apparently similar strings such as *Mary* and *Mary'den*, *Commission* and *Comisión* are related). Both sets of researchers have been experimenting with “light” morphological analysis of the examples to improve the matching process.

This very intuitive approach to translation template generation looks very promising, though one should be aware of certain pitfalls. Let us end this section with a kind of exercise for the reader.<sup>1</sup> Consider the English–Japanese sentence pairs in (23).

- (23) a. The monkey ate a pear ⇒ *Saru wa nashi o tabeta.*  
 b. The man ate a pear ⇒ *Hito wa nashi o tabeta.*

What can be inferred on the basis of these examples alone (answers in a moment)? You should find two lexical pairings, and a possible template. Now if we add a third example, (24), how can you revise your “knowledge” of Japanese?

- (24) The dog ate a rabbit ⇒ *Inu wa usagi o tabeta.*

---

<sup>1</sup> Readers who know Japanese are excused the exercise!

Now here are the answers: from (23) we can correctly infer the lexical pairings *monkey*  $\Leftrightarrow$  *saru* and *man*  $\Leftrightarrow$  *hito*, which leaves us with a template as in (25).

(25) The ... ate a ...  $\Rightarrow$  ... *wa* ... *o tabeta*

The third example (24) appears to suggest two more lexical pairings, *dog*  $\Leftrightarrow$  *inu* and *rabbit*  $\Leftrightarrow$  *usagi*, which happen to be correct. But notice here we have made an assumption that the “slots” in (25) indicated by dots are linked as in (26).

(26) The X ate a Y  $\Rightarrow$  X *wa* Y *o tabeta*

Actually, we do not have any direct evidence of this (another “minimal pair” contrasting *dog* and *rabbit* would do, but we don’t have that data), and we need to be aware of the fact that we have taken something for granted which is not actually in the data. On the basis of our knowledge of how languages generally work, it seems like a fair assumption. But perhaps we can go a step further? On the basis of the examples seen so far, it looks like *wa* might be paired with *the*, and *o* with *a*. Very plausible but, unfortunately not true. In fact *wa* and *o* are subject and object case markers; Japanese does not distinguish definiteness at all. Furthermore, our assumption that the relative order of subject and object is fixed (which lead us to assume the *dog*  $\Leftrightarrow$  *inu* and *rabbit*  $\Leftrightarrow$  *usagi* pairings) actually is not true for Japanese. An example such as (27) would throw our tentative analysis into much turmoil, and shows how careful we have to be in looking at the data and not making too many assumptions based on our linguistic metaknowledge (or, one might say, “prejudice”).

(27) The dog ate the cake  $\Rightarrow$  *Keeki wa inu ga tabeta*

## 6. Conclusion

We have looked at a range of issues related to bilingual parallel corpora. In the context of this workshop on Indian language processing, the conclusion is, one hopes, obvious. None of the techniques described above make any great assumptions of prior knowledge about Indian languages beyond the kind of very general “universal” knowledge about how languages work that all linguists share. Our plan is to collect corpus data in Indian languages and to try to apply some of the techniques described here. We look forward to reporting our results.

## References

- Ahrenberg, L., M. Andersson and M. Merkel. 1998: A simple hybrid aligner for generating lexical correspondences in parallel texts. *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Quebec, pp. 29--35.
- Akiba, Y., M. Ishii, H. Almuallim and S. Kaneda. 1995: Learning English verb selection rules from hand-made rules and translation examples. *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, pp. 206–220.
- Almuallim, H., Y. Akiba, T. Yamazaki, A. Yokoo and S. Kaneda. 1994: Two methods for learning ALT-J/E translation rules from examples and a semantic hierarchy. *COLING 94: The 15th International Conference on Computational Linguistics*, Kyoto, Japan, pp. 57–63.
- Brown, P.F., J.C. Lai and R.L. Mercer. 1991: Aligning sentences in parallel corpora. *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, pp. 169–176.
- Brown, R.D. 1999: Adding linguistic knowledge to a lexical example-based translation system. *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, Chester, England, pp. 22–32.

- Carl, M. 1999: Inducing translation templates for example-based machine translation. *Machine Translation Summit VII*, Singapore, pp. 250–258.
- Chen, S.F. 1993: Aligning sentences in bilingual corpora using lexical information. *31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 9–16.
- Church, K.W. 1993: Char\_align: a program for aligning parallel texts at the character level. *31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 1--8.
- Church, K. and P. Hanks. 1989: Word association norms, mutual information, and lexicography. *27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, British Columbia, pp. 76–83.
- Cicekli, I., and H.A. Güvenir. 1996: Learning translation rules from a bilingual corpus, *NeMLaP-2: Proceedings of the Second International Conference on New Methods in Language Processing*, Ankara, Turkey, pp. 90--97.
- Collins, B. and P. Cunningham. 1995: A methodology for example based machine translation', *CSNLP 1995: 4th Conference on the Cognitive Science of Natural Language Processing*, Dublin, Ireland.
- Dagan, I. 1996: *Bilingual Word Alignment and Lexicon Construction*. Tutorial handout, COLING-96: The 16th International Conference on Computational Linguistics, Copenhagen.
- Dagan, I. and K. Church. 1997: *Termight*: Coordinating humans and machines in bilingual terminology acquisition. *Machine Translation* **12**, 89–107.
- Daille, B. 1995a: Repérage et extraction de terminologie par une approche mixte statistique et linguistique. *Traitements Probabilistes et Corpus, t.a.l.* **36**.1--2, 101–118.
- Daille, B. 1995b: Combined approach for terminology extraction: lexical statistics and linguistic filtering, UCREL Technical Papers, No. 15, Department of Linguistics, Lancaster University.
- Daille, B. 1996: Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In J. Klavans and P. Resnik (eds) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press, Cambridge, Mass., pp.49–66.
- Daille, B., É. Gaussier and J.-M. Langé. 1994: Towards automatic extraction of monolingual and bilingual terminology. *COLING 94: The 15th International Conference on Computational Linguistics*, Kyoto, Japan, pp. 515--521.
- Dunning, T. 1993: Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* **19**, 61–76.
- Fung, P. 1995: Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus. *Proceedings of the Third Workshop on Very Large Corpora*, Cambridge, Mass., pp. 173–183.
- Fung, P. 1998: A statistical view on bilingual lexicon extraction: from parallel corpora to non-parallel corpora. In D. Farwell, L. Gerber and E. Hovy (eds) *Machine Translation and the Information Soup*, Springer, Berlin, pp. 1–17.
- Fung, P. and K.W. Church. 1994: K-vec: a new approach for aligning parallel texts. *COLING 94: The 15th International Conference on Computational Linguistics*, Kyoto, Japan, pp. 1096–1102.
- Fung, P. and K. McKeown. 1997a: A technical word- and term-translation aid using noisy parallel corpora across language groups. *Machine Translation* **12**, 53–87.
- Fung, P. and K. McKeown. 1997b: Finding terminology translations from non-parallel corpora. *Proceedings of the Fifth Workshop on Very Large Corpora*, Beijing and Hong Kong, pp. 192–202.
- Fung, P. and L.Y. Yee. 1998: An IR approach for translating new words from nonparallel, comparable texts. *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Quebec, pp. 414–420.
- Furuse, O. and H. Iida. 1992: An example-based method for transfer-driven machine translation. *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, pp. 139–150.
- Gale, W.A. and K.W. Church. 1991a: A program for aligning sentences in bilingual corpora. *29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, Calif., pp. 177–184.
- Gale, W.A. and K.W. Church. 1991b: Identifying word correspondences in parallel text. *Workshop on Speech and Natural Language Processing*, Asilomar, California.
- Gale, W.A. and K.W. Church. 1991c: Concordances for parallel texts. *Seventh Annual Conference of the UW Centre for New OED and text Research Using Corpora*, Oxford, pp. 40–62.
- Gao, Z-M. 1997: *Automatic Extraction of Translation Equivalents from a Parallel Chinese–English Corpus*, PhD thesis, UMIST, Manchester, England.

- Gaussier, É., D. Hull and S. Ait-Mokhtar. 2000. Term alignment in use: Machine-aided human translation. In J. Véronis (ed.) *Parallel Text Processing: Alignment and Use of Translation Corpora*, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 253–274.
- Gaussier, É. and J.-M. Langé. 1997: Some methods for the extraction of bilingual terminology. In D. Jones and H. Somers (eds) *New Methods in Language Processing*, UL Press, London, pp. 145–153.
- Güvenir, H.A. and I. Cicekli. 1998: Learning translation templates from examples, *Information Systems* **23**, 353–363.
- Güvenir, H.A. and A. Tunç. 1996: Corpus-based learning of generalized parse tree rules for translation. In G. McCalla (ed.) *Advances in Artificial Intelligence*, Berlin: Springer Verlag, pp. 121–132.
- Harris, B. 1988: Bi-text, a new concept in translation theory. *Language Monthly* **54**, 8–10.
- Jain, R., R.M.K. Sinha and A. Jain. 1995: Role of examples in translation. *1995 IEEE International Conference on Systems, Man and Cybernetics*, Vancouver, BC, pp. 1615–1620.
- Johansson, S., J. Ebeling and K. Hofland. 1993: Coding and aligning the English–Norwegian parallel corpus. In K. Ajimer, B. Altenberg and M. Johansson (eds) *Languages in Contrast: A Symposium on Text-Based Cross-Linguistic Studies*, Lund University Press, Lund, pp. 87–112.
- Jones, D. and M. Alexa. 1997: Towards automatically aligning German compounds with English word groups. In D. Jones and H. Somers (eds) *New Methods in Language Processing*, UCL Press, London, pp. 199–206.
- Jones, D.B. and H. Somers. 1995: Bilingual vocabulary estimation from noisy parallel corpora using variable bag estimation. *JADT 1995: III Giornate internazionali di Analisi Statistica dei Dati Testuali*, Rome, Vol. I, pp. 255–262.
- Justeson, J.S. and S.M. Katz. 1995: Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering* **1**, 9–27.
- Kaji, H., Y. Kida and Y. Morimoto. 1992: Learning translation templates from bilingual text. *Proceedings of the fifteenth [sic] International Conference on Computational Linguistics, COLING-92*, Nantes, France, pp. 672–678.
- Kay, M. and M. Röscheisen. 1993: Text translation alignment. *Computational Linguistics* **19**, 121–142.
- Kitamura, M. and Y. Matsumoto. 1996: Automatic extraction of word sequence correspondences in parallel corpora. *Proceedings of the Fourth Workshop on Very Large Corpora*, Copenhagen, Denmark, pp. 79–87.
- Langé, J.M., É. Gaussier and B. Daille. 1997: Bricks and skeletons: some ideas for the near future of MAHT. *Machine Translation* **12**, 39–51.
- Matsumoto, Y. and M. Kitamura. 1995: Acquisition of translation rules from parallel corpora. In R. Mitkov and N. Nicolov (eds), *Recent Advances in Natural Language Processing: Selected Papers from RANLP'95*, Amsterdam (1997): John Benjamins, pp. 405–416.
- McEnery, T. and M. Oakes. 1996: Sentence and word alignment in the CRATER Project. In J. Thomas and M. Short (eds), *Using Corpora for Language Research*, Longman, London, pp. 211–231.
- McEnery, T., J.-M. Langé, M. Oakes and J. Véronis. 1997: The exploitation of multilingual corpora for term extraction. In R. Garside, G. Leech and A. McEnery (eds), *Corpus Annotation: Linguistic Information from Computer Text Corpora*, Addison Wesley Longman, London, pp. 220–230.
- McTait, K., M. Olohan and A. Trujillo. 1999: A building blocks approach to translation memory. *Translating and the Computer 21*, London: Aslib/IMI, [pages not numbered].
- McTait, K. and A. Trujillo. 1999: A language-neutral sparse-data algorithm for extracting translation patterns. *Proceedings of the 8th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 99)*, Chester, England, pp. 98–108.
- Melamed, I.D. 1996a: A geometrical approach to mapping bitext correspondence. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Philadelphia, Pa., pp. 1–12.
- Melamed, I.D. 1996b: Automatic Detection of Omissions in Translation. *COLING-96: The 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, pp. 764–769.
- Nagao, M. 1984: A framework of a mechanical translation between Japanese and English by analogy principle. In A. Elithorn and R. Banerji (eds) *Artificial and Human Intelligence*, North-Holland, Amsterdam, pp. 173–180.
- Nomiyama, H. 1992: Machine translation by case generalization. *Proceedings of the fifteenth [sic] International Conference on Computational Linguistics, COLING-92*, Nantes, France, pp. 714–720.

- Resnik, P. 1998: Parallel strands: a preliminary investigation into mining the web for bilingual text. In D. Farwell, L. Gerber and E. Hovy (eds) *Machine Translation and the Information Soup*, Springer, Berlin, pp. 72--82.
- Resnik, P. 1999: Mining the web for bilingual text. *37th Annual Meeting of the Association for Computational Linguistics*, College Park, Maryland, pp. 527--534.
- Shin, J.H., Y.S. Han and K-S. Choi. 1996: bilingual knowledge acquisition from Korean-English parallel corpus using alignment method (Korean-English alignment at word and phrase level). *COLING-96: The 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, pp. 230--235.
- Simard, M., G. Foster and P. Isabelle. 1992: Using cognates to align sentences in bilingual corpora. *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, pp. 67--82.
- Smadja, F. 1993: Retrieving collocations from text: Xtract. *Computational Linguistics* **19**, 121--142.
- Smadja, F., K.R. McKeown and V. Hatzivassiloglou. 1996: Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* **22**, 1--38.
- Somers, H. 1998: Further experiments in bilingual text alignment. *International Journal of Corpus Linguistics* **3**, 115--150.
- Somers, H. 1999: Review article: Example-based machine translation. *Machine Translation* **14**, 113--158.
- Sperberg-McQueen, C. and L. Burnard. 1994: *Guidelines for electronic text encoding and interchange: TEI-P3, ACH-ACL-ALLC Text Coding Initiative*, Chicago and Oxford.
- Thompson, H.S. 2000. Corpus creation for data-intensive linguistics. In R. Dale, H. Moisl, and H. Somers (eds) *Handbook of Natural Language Processing*, Marcel Dekker, New York, pp. 385--401.
- Wu, D. 2000. Alignment. In R. Dale, H. Moisl and H. Somers (eds) *Handbook of Natural Language Processing*, Marcel Dekker, New York, pp. 415--458.
- Wu, D. and X. Xia. 1994: Learning an English-Chinese lexicon from a parallel corpus. *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, Columbia, Maryland, pp. 206--213.