ALIGNING SENTENCES IN PARALLEL CORPORA

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer

IBM Thomas J. Watson Research Center
P.O. Box 704
Yorktown Heights, NY 10598

ABSTRACT

In this paper we describe a statistical technique for aligning sentences with their translations in two parallel corpora. In addition to certain anchor points that are available in our data, the only information about the sentences that we use for calculating alignments is the number of tokens that they contain. Because we make no use of the lexical details of the sentence, the alignment computation is fast and therefore practical for application to very large collections of text. We have used this technique to align several million sentences in the English-French Hansard corpora and have achieved an accuracy in excess of 99% in a random selected set of 1000 sentence pairs that we checked by hand. We show that even without the benefit of anchor points the correlation between the lengths of aligned sentences is strong enough that we should expect to achieve an accuracy of between 96% and 97%. Thus, the technique may be applicable to a wider variety of texts than we have yet tried.

INTRODUCTION

Recent work by Brown *et al.*, [Brown *et al.*, 1988, Brown *et al.*, 1990] has quickened anew the long dormant idea of using statistical techniques to carry out machine translation from one natural language to another. The lynchpin of their approach is a large collection of pairs of sentences that are mutual translations. Beyond providing grist to the statistical mill, such pairs of sentences are valuable to researchers in bilingual lexicography [Klavans and Tzoukermann, 1990, Warwick and Russell, 1990] and may be useful in other approaches to machine translation [Sadler, 1989].

In this paper, we consider the problem of extracting from parallel French and English corpora pairs sentences that are translations of one another. The task is not trivial because at times a single sentence in one language is translated as two or more sentences in the other language. At other times a sentence, or even a whole passage, may be missing from one or the other of the corpora.

If a person is given two parallel texts and asked to match up the sentences in them, it is natural for him to look at the words in the sentences. Elaborating this intuitively appealing insight, researchers at Xerox and at ISSCO [Kay, 1991, Catizone *et al.*, 1989] have developed alignment algorithms that pair sentences according to the words that they contain. Any such algorithm is necessarily slow and, despite the potential for highly accurate alignment, may be unsuitable for very large collections of text. Our algorithm makes no use of the lexical details of the corpora, but deals only with the number of words in each sentence. Although we have used it only to align parallel French and English corpora from the proceedings of the Canadian Parliament, we expect that our technique would work on other French and English corpora and even on other pairs of languages. The work of Gale and Church , [Gale and Church, 1991], who use a very similar method but measure sentence lengths in characters rather than in words, supports this promise of wider applicability.

THE HANSARD CORPORA

Brown *et al.*, [Brown *et al.*, 1990] describe the process by which the proceedings of the Canadian Parliament are recorded. In Canada, these proceedings are referred to as Hansards.

Our Hansard corpora consist of the Hansards from 1973 through 1986. There are two files for each session of parliament: one English and one French. After converting the obscure text markup language of the raw data to TEX, we combined all of the English files into a single, large English corpus and all of the French files into a single, large French corpus. We then segmented the text of each corpus into tokens and combined the tokens into groups that we call sentences. Generally, these conform to the grade-school notion of a sentence: they begin with a capital letter, contain a verb, and end with some type of sentence-final punctuation. Occasionally, they fall short of this ideal and so each corpus contains a number of sentence fragments and other groupings of words that we nonetheless refer to as sentences. With this broad interpretation, the English corpus contains 85,016,286 tokens in 3,510,744 sentences, and the French corpus contains 97,857,452 tokens in 3,690,425 sentences. The average English sentence has 24.2 tokens, while the average French sentence is about 9.5% longer with 26.5 tokens.

The left-hand side of Figure 1 shows the raw data for a portion of the English corpus, and the right-hand side shows the same portion after we converted it to TEX and divided it up into sentences. The sentence numbers do not advance regularly because we have edited the sample in order to display a variety of phenomena.

In addition to a verbatim record of the proceedings and its translation, the Hansards include session numbers, names of speakers, time stamps, question numbers, and indications of the original language in which each speech was delivered. We retain this auxiliary information in the form of comments sprinkled throughout the text. Each comment has the form $\backslash SCM\{\}$ ... $\backslash ECM\{\}$ as shown on the right-hand side of Figure 1. In addition to these comments, which encode information explicitly present in the data, we inserted *Paragraph* comments as suggested by the space command of which we see an example in the eighth line on the left-hand side of Figure 1.

We mark the beginning of a parliamentary session with a *Document* comment as shown in Sentence 1 on the right-hand side of Figure 1. Usually, when a member addresses the parliament, his name is recorded and we encode it in an *Author* comment. We see an example of this in Sentence 4. If the president speaks, he is referred to in the English corpus as *Mr. Speaker* and in the French corpus as *M. le Président*. If several members speak at once, a shockingly regular occurrence, they are referred to as *Some Hon. Members* in the English and as *Des Voix* in the French. Times are recorded either as exact times on a 24-hour basis as in Sentence 81, or as inexact times of which there are two forms: *Time = Later*, and *Time = Recess*. These are rendered in French as *Time = Plus Tard* and *Time = Recess*. Other types of comments that appear are shown in Table 1.

ALIGNING ANCHOR POINTS

After examining the Hansard corpora, we realized that the comments laced throughout would serve as useful anchor points in any alignment process. We divide the comments into major and minor anchors as follows. The comments *Author = Mr. Speaker*, *Author = M. le Président*, *Author = Some Hon. Members*, and *Author = Des Voix* are called minor anchors. All other comments are called major anchors with the exception of the *Paragraph* comment which is not treated as an anchor at all. The minor anchors are much more common than any particular major anchor, making an alignment based on them less robust against deletions than one based on the major anchors. Therefore, we have carried out the alignment of anchor points in two passes, first aligning the major anchors and then the minor anchors.

Usually, the major anchors appear in both languages. Sometimes, however, through inattention on the part of the translator or other misadventure, the name of a speaker may be garbled or a comment may be omitted. In the first alignment pass, we assign to alignments

```
/*START_COMMENT* Beginning file = 048
101 H002-108 script A *END_COMMENT*/
.TB 029 060 090 099
.PL 060
.LL 120
.NF
The House met at 2 p.m.
.SP
*boMr. Donald MacInnis (Cape Breton
-East Richmond):*ro Mr. Speaker,
I rise on a question of privilege af-
fecting the rights and prerogatives
of parliamentary committees and one
which reflects on the word of two
ministers.
.SP
*boMr. Speaker: *roThe hon. member's
 motion is proposed to the
House under the terms of Standing
Order 43. Is there unanimous consent?
.SP
*boSome hon. Members: *roAgreed.
s*itText*ro)
Question No. 17--*boMr. Mazankowski:
*ro
1. For the period April 1, 1973 to
January 31, 1974, what amount of
money was expended on the operation
and maintenance of the Prime
Minister's residence at Harrington
Lake, Quebec?
.SP
(1415)
s*itLater:*ro)
.SP
*boMr. Cossitt:*ro Mr. Speaker, I rise
on a point of order to ask for
clarification by the parliamentary
secretary.
```

```
 1. \SCM{} Document = 048 101 H002-108
    script A \ECM{}
 2. The House met at 2 p.m.
 3. \SCM{} Paragraph \ECM{}
 4. \SCM{} Author = Mr. Donald MacInnis
    (Cape Breton-East Richmond) \ECM{}
 5. Mr. Speaker, I rise on a question of
    privilege affecting the rights and
    prerogatives of parliamentary
    committees and one which reflects on
    the word of two ministers.
21. \SCM{} Paragraph \ECM{}
22. \SCM{} Author = Mr. Speaker \ECM{}
23. The hon. member's motion is proposed
    to the House under the terms of
    Standing Order 43.
44. Is there unanimous consent?
45. \SCM{} Paragraph \ECM{}
46. \SCM{} Author = Some hon. Members
    \ECM{}
47. Agreed.
61. \SCM{} Source = Text \ECM{}
62. \SCM{} Question = 17 \ECM{}
63. \SCM{} Author = Mr. Mazankowski
    \ECM{}
64. 1.
65. For the period April 1, 1973 to
    January 31, 1974, what amount of
    money was expended on the operation
    and maintenance of the Prime
    Minister's residence at Harrington
    Lake, Quebec?
66. \SCM{} Paragraph \ECM{}
81. \SCM{} Time = (1415) \ECM{}
82. \SCM{} Time = Later \ECM{}
83. \SCM{} Paragraph \ECM{}
84. \SCM{} Author = Mr. Cossitt \ECM{}
85. Mr. Speaker, I rise on a point of
    order to ask for clarification by
    the parliamentary secretary.
```

Figure 1: A sample of text before and after cleanup

a cost that favors exact matches and penalizes omissions or garbled matches. Thus, for example, we assign a cost of 0 to the pair *Time = Later* and *Time = Plus Tard,* but a cost of 10 to the pair *Time = Later* and *Author = Mr. Bateman.* We set the cost of a deletion at 5. For two names, we set the cost by counting the number of insertions, deletions, and substitutions necessary to transform one name, letter by letter, into the other. This value is then reduced to the range 0 to 10.

Given the costs described above, it is a standard problem in dynamic programming to find that alignment of the major anchors in the two corpora with the least total cost [Bellman, 1957]. In theory, the time and space required to find this alignment grow as the product of the lengths of the two sequences to be aligned. In practice, however, by using thresholds and the partial traceback technique described by Brown, Spohrer, Hochschild, and Baker , [Brown *et al.,* 1982], the time required can be made linear in the length of the sequences, and the space can be made constant. Even so, the computational demand is severe since, in places, the two corpora are out of alignment by as many as 90,000 sentences owing to mislabelled or missing files.

This first pass renders the data as a sequence of sections between aligned major anchors. In the second pass, we accept or reject each section in turn according to the population of minor anchors that it contains. Specifically, we accept a section provided that, within the section, both corpora contain the same number of minor anchors in the same order. Otherwise, we reject the section. Altogether, we reject about 10% of the data in each corpus. The minor anchors serve to divide the remaining sections into subsections that range in size from one sentence to several thousand sentences and average about ten sentences.

ALIGNING SENTENCES AND PARAGRAPH BOUNDARIES

We turn now to the question of aligning the individual sentences in a subsection between minor anchors. Since the number of sentences in the French corpus differs from the number in the English corpus, it is clear that they cannot be in one-to-one correspondence throughout. Visual inspection of the two corpora quickly reveals that although roughly 90% of the English sentences correspond to single French sentences, there are many instances where a single sentence in one corpus is represented by two consecutive sentences in the other. Rarer, but still present, are examples of sentences that appear in one corpus but leave no trace in the other. If one is moderately well acquainted with both English and French, it is a simple matter to decide how the sentences should be aligned. Unfortunately, the sizes of our corpora make it impractical for us to obtain a complete set of alignments by hand. Rather, we must necessarily employ some automatic scheme.

English	**French**
Source = English	Source = Traduction
Source = Translation	Source = Francais
Source = Text	Source = Texte
Source = List Item	Source = List Item
Source = Question	Source = Question
Source = Answer	Source = Reponse

Table 1: Examples of comments

It is not surprising and further inspection verifies that the number of tokens in sentences that are translations of one another are correlated. We looked, therefore, at the possibility of obtaining alignments solely on the basis of sentence lengths in tokens. From this point of view, each corpus is simply a sequence of sentence lengths punctuated by occasional paragraph markers. Figure 2 shows the initial portion of such a pair of corpora. We have circled groups of sentence lengths to show the correct alignment. We call each of the groupings a *bead.* In this example, we have an *ef*-bead followed by an *eff*-bead followed by an *e*-bead followed by a $\P_e\P_f$-bead. An alignment, then, is simply a sequence of beads that accounts for the observed sequences of sentence lengths and paragraph markers. We imagine the sentences in a subsection to have been generated by a pair of random processes, the first pro-

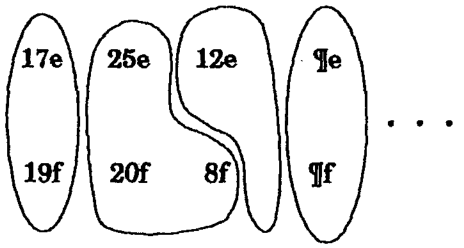

Figure 2: Division of aligned corpora into beads

Bead	*Text*
e	one English sentence
f	one French sentence
ef	one English and one French sentence
eef	two English and one French sentence
eff	one English and two French sentences
$¶_e$	one English paragraph
$¶_f$	one French paragraph
$¶_e¶_f$	one English and one French paragraph

Table 2: Alignment Beads

ducing a sequence of beads and the second choosing the lengths of the sentences in each bead.

Figure 3 shows the two-state Markov model that we use for generating beads. We assume that a single sentence in one language lines up with zero, one, or two sentences in the other and that paragraph markers may be deleted. Thus, we allow any of the eight beads shown in Table 2. We also assume that $\Pr(e) = \Pr(f)$, $\Pr(eff) = \Pr(eef)$, and $\Pr(¶_e) = \Pr(¶_f)$.

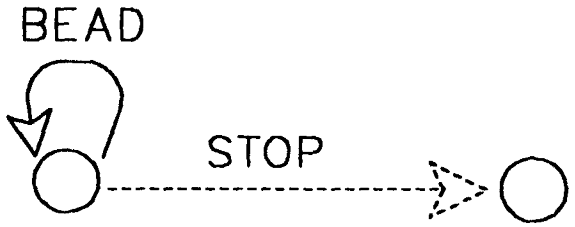

Figure 3: Finite state model for generating beads

Given a bead, we determine the lengths of the sentences it contains as follows. We assume the probability of an English sentence of length ℓ_e given an e-bead to be the same as the probability of an English sentence of length ℓ_e in the text as a whole. We denote this probability by $\Pr(\ell_e)$. Similarly, we assume the probability of a French sentence of length ℓ_f given an f-bead to be $\Pr(\ell_f)$. For an ef-bead, we assume that the English sentence has length ℓ_e with probability $\Pr(\ell_e)$ and that log of the ratio of length of the French sentence to the length of the English sentence is normally distributed with mean μ and variance σ^2. Thus, if $r = \log(\ell_f/\ell_e)$, we assume that

$$\Pr(\ell_f|\ell_e) = \alpha \exp[-(r-\mu)^2/(2\sigma^2)], \qquad (1)$$

with α chosen so that the sum of $\Pr(\ell_f|\ell_e)$ over positive values of ℓ_f is equal to unity. For an eef-bead, we assume that each of the English sentences is drawn independently from the distribution $\Pr(\ell_e)$ and that the log of the ratio of the length of the French sentence to the sum of the lengths of the English sentences is normally distributed with the same mean and variance as for an ef-bead. Finally, for an eff-bead, we assume that the length of the English sentence is drawn from the distribution $\Pr(\ell_e)$ and that the log of the ratio of the sum of the lengths of the French sentences to the length of the English sentence is normally distributed as before. Then, given the sum of the lengths of the French sentences, we assume that the probability of a particular pair of lengths, ℓ_{f_1} and ℓ_{f_2}, is proportional to $\Pr(\ell_{f_1})\Pr(\ell_{f_2})$.

Together, these two random processes form a hidden Markov model [Baum, 1972] for the generation of aligned pairs of corpora. We determined the distributions, $\Pr(\ell_e)$ and $\Pr(\ell_f)$, from the relative frequencies of various sentence lengths in our data. Figure 4 shows for each language a histogram of these for sentences with fewer than 81 tokens. Except for lengths 2 and 4, which include a large number of formulaic sentences in both the French and the English, the distributions are very smooth.

For short sentences, the relative frequency is a reliable estimate of the corresponding probability since for both French and English we have more than 100 sentences of each length less than 81. We estimated the probabilities

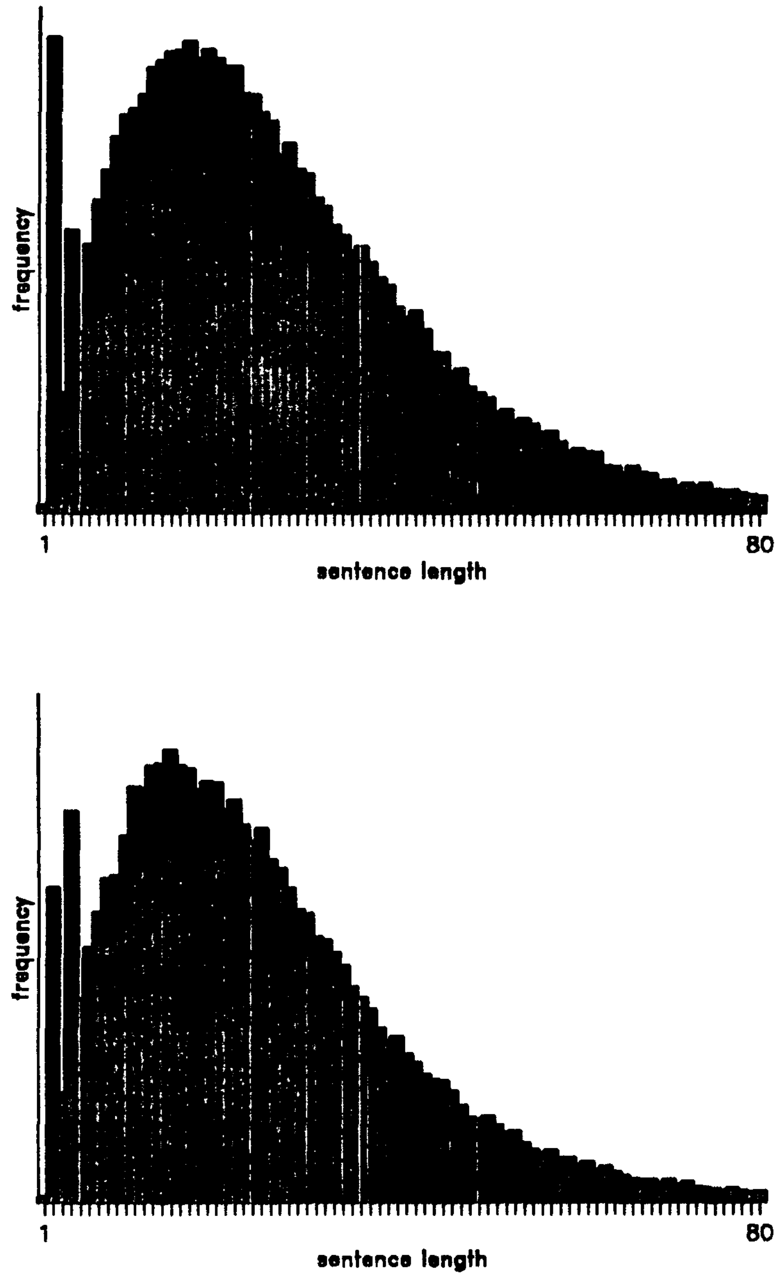

Figure 4: Histograms of French (top) and English (bottom) sentence lengths

of greater lengths by fitting the observed frequencies of longer sentences to the tail of a Poisson distribution.

We determined all of the other parameters by applying the EM algorithm to a large sample of text [Baum, 1972, Dempster *et al.*, 1977]. The resulting values are shown in Table 3. From these parameters, we can see that 91% of the English sentences and 98% of the English paragraph markers line up one-to-one with their French counterparts. A random variable x, the log of which is normally distributed with mean μ and variance σ^2, has mean value $\exp(\mu + \sigma^2/2)$. We can also see, therefore, that the total length of the French text in an *ef*-, *eef*-, or *eff*-bead should be about 9.8% greater on average than the total length of the corresponding English text. Since most sentences belong to *ef*-beads, this is close to the value of 9.5% given in Section 2 for the amount by which the length of the average French sentences exceeds that of the average English sentence.

Parameter	*Estimate*
$\Pr(e), \Pr(f)$	.007
$\Pr(ef)$	.690
$\Pr(eef), \Pr(eff)$	.020
$\Pr(¶_e), \Pr(¶_f)$	.005
$\Pr(¶_e ¶_f)$	.245
μ	.072
σ^2	.043

Table 3: Parameter estimates

We can compute from the parameters in Table 3 that the entropy of the bead production process is 1.26 bits per sentence. Using the parameters μ and σ^2, we can combine the observed distribution of English sentence lengths shown in Figure 4 with the conditional distribution of French sentence lengths given English sentence lengths in Equation (1) to obtain the joint distribution of French and English sentences lengths in *ef*-, *eef*-, and *eff*-beads. From this joint distribution, we can compute that the mutual information between French and English sentence lengths in these beads is 1.85 bits per sentence. We see therefore that, even in the absence of the anchor points produced by the first two passes, the correlation in sentence lengths is strong enough to allow alignment with an error rate that is asymptotically less than 100%. Heartening though such a result may be to the theoretician, this is a sufficiently coarse bound on the error rate to warrant further study. Accordingly, we wrote a program to simulate the alignment process that we had in mind. Using $\Pr(\ell_e)$, $\Pr(\ell_f)$, and the parameters from Table 3, we generated an artificial pair of aligned corpora. We then determined the most probable alignment for the data. We recorded the fraction of *ef*-beads in the most probable alignment that did not correspond to *ef*-beads in the true alignment as the error rate for the process. We repeated this process many thousands of times and found that we could expect an error rate of about 0.9% given the frequency of anchor points from the first two passes.

By varying the parameters of the hidden Markov model, we explored the effect of anchor points and paragraph markers on the accuracy of alignment. We found that with paragraph markers but no anchor points, we could expect an error rate of 2.0%, with anchor points but no paragraph markers, we could expect an error rate of 2.3%, and with neither anchor points nor paragraph markers, we could expect an error rate of 3.2%. Thus, while anchor points and paragraph markers are important, alignment is still feasible without them. This is promising since it suggests that one may be able to apply the same technique to data where frequent anchor points are not available.

RESULTS

We applied the alignment algorithm of Sections 3 and 4 to the Canadian Hansard data described in Section 2. The job ran for 10 days on an IBM Model 3090 mainframe under an operating system that permitted access to 16 megabytes of virtual memory. The most probable alignment contained 2,869,041 *ef*-beads. Some of our colleagues helped us

And love and kisses to you, too.	Pareillement.
... mugwumps who sit on the fence with their mugs on one side and their wumps on the other side and do not know which side to come down on.	... en voulant ménager la chèvre et le choux ils n'arrivent pas à prendre parti.
At first reading, she may have.	Elle semble en effet avoir un grief tout a fait valable, du moins au premier abord.

Table 4: Unusual but correct alignments

examine a random sample of 1000 of these beads, and we found 6 in which sentences were not translations of one another. This is consistent with the expected error rate of 0.9% mentioned above. In some cases, the algorithm correctly aligns sentences with very different lengths. Table 4 shows some interesting examples of this.

REFERENCES

[Baum, 1972] Baum, L. (1972). An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities*, 3:1–8.

[Bellman, 1957] Bellman, R. (1957). *Dynamic Programming.* Princeton University Press, Princeton N.J.

[Brown *et al.*, 1982] Brown, P., Spohrer, J., Hochschild, P., and Baker, J. (1982). Partial traceback and dynamic programming. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1629–1632, Paris, France.

[Brown *et al.*, 1990] Brown, P. F., Cocke, J., DellaPietra, S. A., DellaPietra, V. J., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79-85.

[Brown *et al.*, 1988] Brown, P. F., Cocke, J., DellaPietra, S. A., DellaPietra, V. J., Jelinek, F., Mercer, R. L., and Roossin, P. S. (1988). A statistical approach to language translation. In *Proceedings of the 12th International Conference on Computational Linguistics*, Budapest, Hungary.

[Catizone *et al.*, 1989] Catizone, R., Russell, G., and Warwick, S. (1989). Deriving translation data from bilingual texts. In *Proceedings of the First International Acquisition Workshop*, Detroit, Michigan.

[Dempster *et al.*, 1977] Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38.

[Gale and Church, 1991] Gale, W. A. and Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California.

[Kay, 1991] Kay, M. (1991). Text-translation alignment. In *ACH/ALLC '91: "Making Connections" Conference Handbook*, Tempe, Arizona.

[Klavans and Tzoukermann, 1990] Klavans, J. and Tzoukermann, E. (1990). The bicord system. In *COLING-90*, pages 174–179, Helsinki, Finland.

[Sadler, 1989] Sadler, V. (1989). *The Bilingual Knowledge Bank - A New Conceptual Basis for MT.* BSO/Research, Utrecht.

[Warwick and Russell, 1990] Warwick, S. and Russell, G. (1990). Bilingual concordancing and bilingual lexicography. In *EURALEX 4th International Congress*, Málaga, Spain.