Concordances for Parallel Text

Kenneth W. Church
William A. Gale

1. Introduction

Researchers in both machine translation (e.g., Brown *et al*, 1990) and bilingual lexicography (e.g., Klavans and Tzoukermann, 1990a, 1990b) have recently become interested in studying *parallel texts* (also known as *bilingual corpora*), bodies of text such as the Canadian Hansards (parliamentary debates) which are available in multiple languages (such as French and English). Much of the current excitement surrounding parallel texts was initiated by Brown *et al.* (1990), who outline a self-organizing method for using these parallel texts to build a machine translation system. This approach is extremely controversial among researchers in machine translation because it questions many of the basic assumptions that have dominated the field since the 1950s when Chomsky and others pointed out that statistical ngram methods are incapable of modeling certain syntactic constraints such as agreement over long distances (Chomsky, 1957, pp. 15-17). It is possible that the Brown *et al.* approach is deeply flawed for many of the reasons that were discussed in the 1950s. Nevertheless, there is a growing community of researchers in corpus-based linguistics who are becoming convinced that the approach is worth pursuing because there is a very good chance that it will produce a number of lexical resources that could be of great value to their research.

In particular, this paper will discuss a range of bilingual concordancing tools that have been implemented. At the last Waterloo conference, there were a few objections to the use of parallel texts, suggesting that monolingual corpora are easier to obtain, and solve most of the needs of bilingual lexicographers. In fact, we suspect the real objection might have been to the hypothetical nature of the proposals that were being discussed last year. Much of the discussion surrounded suggestions for how parallel texts might be aligned. Now that it has been done and we are eager to make bilingual concordances available upon request, we hope that the time has come for bilingual lexicographers to start using these tools on a daily basis. We also argue that these tools should be used for monolingual applications in addition to bilingual applications.

Unless stated otherwise, all of the results reported here were computed over our corpus of Canadian Hansards (1986-1989), which is being distributed by the ACL/ DCI.

2. Sentence Alignment

2.1 Problem Statement

Following Brown *et al.* (1990), we begin by aligning the parallel texts at the sentence level. That is, we input a pair of texts and output a sequence of aligned regions as illustrated in Table 1. (The text used in this example was extracted from the UBS (Union Bank of Switzerland) corpus, which is available from the ACL/ DCI.) The problem for the alignment program

is deciding where to place the horizontal lines. In our experience, 90% of the English sentences match exactly one French sentence, but other possibilities, especially two sentences matching one (2-1) or one matching two (1-2), are not uncommon. Table 1 illustrates a particularly hard paragraph; only two of the sentences line up one for one.

Table 1: Output from Alignment Program

English	French
According to our survey, 1988 sales of mineral water and soft drinks were much higher than in 1987, reflecting the growing popularity of these products. Cola drink manufacturers in particular achieved above-average growth rates.	Quant aux eaux mińerales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement suṕerieures`a celles de 1987, pour les boissons`a base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.	La progression des chiffres d'affaires ŕesulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.	L'emploi et les investissements ont ́egalement augmenté.
Following a two-year transitional period, the new Foodstuffs Ordinance for Mineral Water came into effect on April 1, 1988. Specifically, it contains more stringent requirements regarding quality consistency and purity guarantees.	La nouvelle ordonnance fed́erale sur les denŕees alimentaires concernant entre autres les eaux mińerales, entŕee en vigueur le 1er avril 1988 apr`es une ṕeriode transitoire de deux ans, exige surtout une plus grande constance dans la qualiŕe et une garantie de la pureté.

2.2 Previous Work

There has been quite a bit of recent work on sentence alignment, e.g., (Kay and Ṛoscheisen, 1988), (Catizone, Russell, and Warwick, to appear), (Brown, Lai and Mercer, 1991) and (Gale and Church, 1991a). The details of these alignment programs are probably not very important to the main point of this paper. The important point is that large quantities of aligned text are now available (from the ACL/ DCI), and there should be plenty more in the future now that a number of places have been able to replicate some of these programs. If the reader is interested in replicating (Gale and Church, 1991a), we suggest that he or she ask us for a copy of the unabridged version, and pay special attention to the c-code in the appendix.

All four methods divide the two texts into sentences and then use dynamic programming to join sentences into regions as necessary subject to a local distance measure. (Kay and Ṛoscheisen, 1988) and (Catizone, Russell, and Warwick, to appear) use a distance measure based on ''corresponding'' words such as *house* and *maison*. Unfortunately, the task of determining word correspondences is computationally expensive as we will see and consequently, this approach has not yet been scaled up to deal with a large text such as a few years of the Canadian Hansards.

(Brown, Lai and Mercer, 1991) and (Gale and Church, 1991a) use a weaker method that is easier to scale up; the distance measure ignores the words and just counts the length of the two regions. It turns out that the length of a text is highly correlated with the length of its translation. That is, a longer sentence will be translated into a longer sentence (or perhaps two sentences whose total length is not very different from the original sentence), whereas a shorter sentence will be translated into a shorter sentence. The two approaches differ in how they count the length of a region. (Brown, Lai and Mercer, 1991) count words whereas (Gale and Church, 1991a) count characters. In our experience, characters seem to work slightly better because longer words tend to be translated into longer words (or several shorter words), whereas shorter words tend to be translated into shorter words.

The details of the alignment program are described in the next two sections. It is not particularly important that the reader be able to follow this discussion unless he or she is thinking of trying to replicate the procedure. The important point is that it is now possible to align sentences with sufficient accuracy to be extremely useful for constructing bilingual concordances.

2.3 Distance Measure

Our distance measure, *d*, is an estimate of $-log\ Pr(match|\delta)$, where δ encodes what we know about the lengths of the two regions in question, l_1 and l_2. The log is introduced here so that adding distances will produce desirable results. The distance measure is based on the assumption that each character in one language L_1 gives rise to a random number of characters in the other language L_2. We assume these random variables are independent and identically distributed with a normal distribution. The model is then specified by the mean, *c*, and the variance, s^2, of this distribution. *c* is the expected number of characters in L_2 per character in L_1, and s^2 is the variance of the number of characters in L_2 per character in L_1. The parameters *c* and s^2 were determined empirically to be 1 and 6.8, respectively. One could estimate different values for these parameters for each particular pair of source and target languages, but we have not found that performance improves by doing so.

We now appeal to Bayes Theorem to estimate $Pr(match|\delta)$ as $Pr(\delta|match)\ Pr(match)$. The first factor is estimated by

$$Pr(\delta|match) = 2\ (1 - Pr(|\delta|))$$

where $Pr(|\delta|)$ is the probability that a random variable, *z*, with a standardized (mean zero, variance one) normal distribution, has a magnitude at least as large as $|\delta|$ Many statistics textbooks provide a table for computing this. The program computes δ directly from the lengths of the two portions of text, l_1 and l_2, using the formula:

$$\delta = \frac{l_2 - l_1\ c}{\sqrt{l_1\ s^2}}$$

and then computes $Pr(|\delta|)$ by integrating a standard normal distribution. This completes the description of the first factor.

The second factor, $Pr(match)$, is fit with the values in Table 2 (below), which were determined empirically. We have found that a sentence in one language normally matches exactly one sentence in the other language (1-1), but other possibilities are not uncommon.

Table 2: Pr(match)

Category	Frequency	Pr(match)
1-1	1167	0.89
1-0 or 0-1	13	0.0099
2-1 or 1-2	117	0.089
2-2	15	0.011
total	1312	1.00

The distance function *d* is defined in a general way to allow for insertions, deletion, substitution, etc. The function takes four arguments: x_1, y_1, x_2, y_2.

1. Let $d(x_1, y_1; 0, 0)$ be the cost of substituting x_1 with y_1,
2. $d(x_1, 0; 0, 0)$ be the cost of deleting x_1,
3. $d(0, y_1; 0, 0)$ be the cost of insertion of y_1,
4. $d(x_1, y_1; x_2, 0)$ be the cost of contracting x_1 and x_2 to y_1,
5. $d(x_1, y_1; 0, y_2)$ be the cost of expanding x_1 to y_1 and y_2, and
6. $d(x_1, y_1; x_2, y_2)$ be the cost of merging x_1 and x_2 and matching with y_1 and y_2.

2.4 The Dynamic Programming Algorithm

The algorithm is summarized in the following recursion equation. Let s_i, $i = 1 \cdots I$, be the sentences of one language, and t_j, $j = 1 \cdots J$, be the translations of those sentences in the other language. Let d be the distance function described in the previous section, and let $D(i,j)$ be the minimum distance between sentences $s_{1, \cdots i}$ and their translations $t_{1, \cdots j}$, under the maximum likelihood alignment. $D(i,j)$ is computed by minimizing over six cases (substitution, deletion, insertion, contraction, expansion and merger) which, in effect, impose a set of slope constraints. That is, $D(i, j)$ is defined by the following recursion with the initial condition $D(i, j) = 0$.

$$D(i, j) = \min \begin{cases} D(i, j-1) + d(0, t_j; 0, 0) \\ D(i-1, j) + d(s_i, 0; 0, 0) \\ D(i-1, j-1) + d(s_i, t_j; 0, 0) \\ D(i-1, j-2) + d(s_i, t_j; 0, t_{j-1}) \\ D(i-2, j-1) + d(s_i, t_j; s_{i-1}, 0) \\ D(i-2, j-2) + d(s_i, t_j; s_{i-1}, t_{j-1}) \end{cases}$$

2.5 Evaluation

All of the UBS sentences were aligned by a primary judge, a native speaker of English with a reading knowledge of French and German. Two additional judges, a native speaker of French and a native speaker of German, respectively, were used to check the primary judge on 43 of the more difficult paragraphs (230 sentences). Both of the additional judges were also fluent in English, having spent the last few years living and working in the United States, though they were both more comfortable with their native language than with English.

The materials were prepared in order to make the task somewhat less tedious for the judges. Each paragraph was printed in three columns, one for each of the three languages: English, French and German. Blank lines were inserted between sentences. The judges were asked to draw lines between matching sentences. The judges were also permitted to draw a line between a sentence and ‘‘null’’ if they thought that the sentence was not translated. For the purposes of this evaluation, two sentences were defined to ‘‘match’’ if they shared a common clause. (In a few cases, a pair of sentences shared only a phrase or a word, rather than a clause; these sentences did not count as a ‘‘match’’ for the purposes of this experiment.)

After checking the primary judge with the other two judges, it was decided that the primary judge’s results were sufficiently reliable that they could be used as a standard for evaluating the program. The primary judge made only two mistakes on the 43 hard paragraphs (one French mistake and one German mistake), whereas the program made 44 errors on the same materials. Since the primary judge’s error rate is so much lower than that of the program, it was decided that we needn’t be concerned with the primary judge’s error rate. If the program and the judge disagree, we can assume that the program is probably wrong.

The 43 ‘‘hard’’ paragraphs were selected by looking for sentences that mapped to something other than themselves after going through both German and French. Specifically, for each English sentence, we

attempted to find the corresponding German sentences, and then for each of them, we attempted to find the corresponding French sentences, and then we attempted to find the corresponding English sentences, which should hopefully get us back to where we started. The 43 paragraphs included all sentences in which this process could not be completed around the loop. This relatively small group of paragraphs (23% of all paragraphs) contained a relatively large fraction of the program's errors (82%). Thus, there seems to be some verification that this trilingual criterion does in fact succeed in distinguishing more difficult paragraphs from less difficult ones.

There are three pairs of languages: English-German, English-French and French-German. We will report on just the first two. (The third pair is probably very dependent on the first two.) Errors are reported with respect to the judge's responses. That is, for each of the ''matches'' that the primary judge found, we report the program as correct if it found the ''match'' and incorrect if it didn't. This convention allows us to compare performance across different algorithms in a straightforward fashion.

The program made 36 errors out of 621 total alignments (5.8%) for English-French, and 19 errors out of 695 (2.7%) alignments for English-German. Overall, there were 55 errors out of a total of 1316 alignments (4.2%). We also found that the distance measure is a good predictor of performance. That is, when the program thinks that it is doing badly, it probably is. Actually, this is a desirable property because it makes it possible to extract a large subcorpus which has a much smaller error rate. By selecting the best scoring 80% of the alignments, the error rate can be reduced from 4% to 0.7%.

Less formal tests of the error rate in the Hansards suggest that the overall error rate is about 2%, while the error rate for the easy 80% of the sentences is about 0.4%. Apparently the Hansard translations are more literal than the UBS reports. It took 20 hours of real time on a Sun 4 to align 367 days of Hansards, or 3.3 minutes per Hansard-day. The 367 days of Hansards contain about 890,000 sentences or about 37 million ''words'' (tokens). About half of the computer time is spent identifying tokens, sentences, and paragraphs, and about half of the time is spent in the *align* program itself.

3. Concordances Based on Aligned Sentences

After the sentences have been aligned, it is possible to provide the lexicographer with bilingual concordances. Table 3, for example, shows a small sample of a bilingual concordance contrasting the uses of *drug* that are translated as *médicaments* with those that are translated as *drogue*. There were actually 1023 instances of *drug*/ *médicaments* and 145 instances of *drug*/ *drogue* in our sample Canadian Hansards (18 million words of English and 18.8 million words of French).

The sentence-based concordance tool makes use of a pre-computed index indicating which sentences in the one language correspond to which sentences in the other language. The program takes two words as input: (1) a probe in one language (e.g., the English word *risk*) and (2) a translation of the probe in the other language (e.g., *risque*). The program searches for pairs of sentences and their translations that contain the probe and its translation.

We have found the sentence-based concordance tool to be useful for studing sense-disambiguation questions, a major open problem in computational linguistics. We suspect, however, that the tool might also help lexicographers, though we are just beginning to exploring these possibilities. If there are any lexicographers who might be interested in studing these questions in more detail, they should certainly contact us and see what the possibilities are for joint collaboration.

The *drug* example is a good one for illustrating how bilingual concordances can be used to study sense-disambiguation questions. The example shows very nicely how the French translation can be used to disambiguate the English text, at least in some cases. Such a sense-disambiguated corpus would be a valuable resource for all kinds of natural language applications. In particular, the corpus could be used to develop and test sense-disambiguation algorithms. For example, if you have an algorithm that is intended to distinguish the ''good'' sense of *drug* from the ''bad'' sense of *drug*, then you might apply your algorithm to all of the uses of *drug* in the English portion of the parallel corpus and use the French text to grade the results. That is, you would say that your program was correct if it identified a use of *drug* as a ''good'' sense and it was translated as *médicament*, and you would say that the program was incorrect if the

Table 3: A Bilingual Concordance Based on Aligned Sentences

drug/ḿedicaments (‘‘good’’ drug)

208648: d states in the form of changes to n these patent	drug laws during or prior to the free trade negoti
18261978: on aux ét́ats-unis en modifiant ces lois sur les	ḿedicaments breveté́es , pendant ou avant les ne
660714: s of our lumber to the united states . SENT %s.c.	drug patent law amendments SENT %brev SENT > | mr
18731544: . les modifications à la loi sur les brevets de	ḿedicaments SENT %brev SENT > | m . steven w . l
894609: mely the human body , so that it resists whatever	drug was used to fight it . SENT there is ample pr
18974531: de sorte qu’ il ŕesiste maintenant à tous les	ḿedicaments . SENT nous avons maintenant la preu
952490: trade talks and a possible canadian move to hike	drug prices in canada when documents released yest
19035379: atude notamment sur le rapport entre les prix des	ḿedicaments et le libre-échange ? SENT > | des
952520: e the relationship between , among other things ,	drug prices and free trade ? SENT > | some hon . m
19035379: atude notamment sur le rapport entre les prix des	ḿedicaments et le libre-échange ? SENT > | des
1106045: the americans a commitment to bring in a generic	drug bill which will increase the drug costs of ca
19194437: ricains de pŕesenter un projet de loi sur les	ḿedicaments non breveté́es qui fera perdre 200 o

drug/drogue (‘‘bad’’ drug)

941162: ted that the chief promoters of that program were	drug smugglers and gun-runners and in no way were
19023479: eurs de ce programme étaient des trafiquants de	drogue et d’ armes , parfaitement incapables de fa
1288675: ned out that these people were in fact a group of	drug smugglers and gun runners who knew nothing ab
19384390: taient en ŕealité un groupe de trafiquants de	drogue et de marchands d’ armes qui ne connaissaie
1525398: r . speaker , is the minister suggesting that the	drug traffickers , offenders and those with armed
19630303: order le statut de ŕefugíe aux trafiquants de	drogue , aux criminels et aux personnes arḿees ?
1808214: al with problems they might have with alcohol and	drug abuse , as well as generally adjusting to com
19924350: nt ŕegler leurs probl̀emes d’ alcoolisme et de	drogue et c’ est à ce milieu enfin qu’ ils devro
2224048: my riding was one which helped young people with	drug problems . SENT that was cut back altogether
20355996: roupes aidait les jeunes qui s’ adonnaient à la	drogue . SENT il ne touche plus un sou . SENT > si

program identified the use as a ‘‘good’’ sense and it was translated as *drogue*.

Thus, the availability of the French text provides a valuable research opportunity, for both monolingual and bilingual applications. The French text can be used to help clarify distinctions in the English text that may not be obvious to a dumb computer. Brown *et al.* (1991b) has explored this possibility in some detail. Their approach takes advantage of the same sorts of clues that a lexicographer would be interested in. Namely, the ‘‘good’’ sense of drug is associated with words like *prices* and *prescription* whereas the bad sense is associated with words like *abuse* and *paraphernalia*.

4. An Information Retrieval (IR) Approach to Sense Disambiguation

We have been experimenting with an information retrieval approach to sense disambiguation. In the training phase, we collect a number of instances of *drug* that are translated as *méd̂icament*, and a number of instances of *drug* uses that are translated as *drogues*. Then in the testing phase, we are given a new instance of *drug*, and are asked to assign the instance to one of the two senses. We attempt to answer this question by comparing the context of the unknown instance with contexts of known instances.

Basically we are treating contexts as analogous to documents in an information retrieval setting. Just as the probabilistic retrieval model (Salton, 1989, section 10.3) sorts documents *d* by

$$score(d) = \prod_{token \text{ in } d} \frac{Pr(token|rel)}{Pr(token|irrel)}$$

we will sort contexts *c* by

$$score(c) = \prod_{token \text{ in } c} \frac{Pr(token|sense_1)}{Pr(token|sense_2)}$$

where $Pr(token|sense)$ is an estimate of the probability that *token* appears in the context of *sense*. Contexts are defined arbitrarily as a window 50 words to the left and 50 words to the right of the probe.

The performance on *drug* and *sentence* are very encouraging (94% and 95%, respectively). We should mention, however, that other words may not work as well as these. So far, we have only looked at a few words; we intend to report more fully on this sense-disambiguation algorithm after we have had a chance to run some more experiments.

These words may be particularly well-suited for this sense-disambiguation procedure because their contexts often contain some very strong clues such as *prices* and *abuse*. Table 4 below shows some words that have a large value for

$$freq(token, sense_1) \log \frac{Pr(token|sense_1)}{Pr(token|sense_2)}$$

These words tend to play an important role in the scoring of contexts. Note that many of the words in Table 4 are, in fact, highly associated with the sense that they are listed under. This suggests that the statistics are doing something fairly reasonable.[1]

Table 4: Contextual Clues for Sense Disambiguation

Word	Sense	Contextual Clues
drug	medicaments	prices, prescription, patent, increase, generic, companies, upon, consumers, higher, price, consumer, multinational, pharmaceutical, costs
drug	drogues	abuse, paraphernalia, illicit, use, trafficking, problem, food, sale, alcohol, shops, crime, cocaine, epidemic, national, narcotic, strategy, head, control, marijuana, welfare, illegal, traffickers, controlled, fight, dogs
sentence	peine	inmate, parole, serving, a, released, prison, mandatory, judge, after, years, who, death, his, murder
sentence	phrase	I, read, second, amended, ‘‘, ’’, protects, version, just, letter, quote, word, ..., last, amendment, insults, assures, quotation, first

1. Accents have been removed from the words in this table.

5. Bar-Hillel's Argument

It would be a major breakthrough if the availability of parallel text made it possible to make progress on the sense-disambiguation problem. Sense disambiguation has been a major obstacle blocking progress in machine translation as well as many other areas of computational linguistics for a long time. Bar-Hillel, who had been one of the leaders in machine translation, abandoned the field when he couldn't see how a program could disambiguate the word *pen* in the very simple English discourse:

> Little John was looking for his toy box.
> Finally he found it.
> *The box was in the pen.*
> John was very happy.

Bar-Hillel (1960, p. 159) argued that:

> ''Assume, for simplicity's sake, that *pen* in English has only the following two meanings: (1) a certain writing utensil, (2) an enclosure where small children can play. I now claim that no existing or imaginable program will enable an electronic computer to determine that the word *pen* in the given sentence within the given context has the second of the above meanings, whereas every reader with a sufficient knowledge of English will do this 'automatically.' ''

Bar-Hillel realized that there was no limit to the number of facts that a computer would need to know in order to disambiguate simple words like *pen*.

> ''Whenever I offered this argument to one of my colleagues working on MT, their first reaction was: 'But why not envisage a system which will put this knowledge at the disposal of the translation machine?' Understandable as this reaction is, it is very easy to show its futility. What such a suggestion amounts to, if taken seriously, is the requirement that a translation machine should not only be supplied with a dictionary but also with a universal encyclopedia. This is surely utterly chimerical and hardly deserves any further discussion. Since, however, the idea of a machine with encyclopedic knowledge has popped up also on other occasions, let me add a few words on this topic. The number of facts we human beings know is, in a certain very pregnant sense, infinite.''

Everyone would agree that the lexicon contains a large number of facts. It will require a monumental effort, perhaps larger than Murray's OED project, to stuff these facts into a computer. Perhaps the task is hopeless, as Bar-Hillel suggests. But we prefer to view these same facts in a more optimistic light. The fact that French and English are as different as they are makes for a valuable research opportunity, as mentioned above. The French text can be used to clarify distinctions in the English that may not be obvious to a dumb computer. Thus, we can use the French text to disambiguate word-senses in the English, producing a large sense-disambiguated corpus to develop and test word-sense disambiguation algorithms such as the one discussed in the previous section.

The advantage of the statistical approach is that it can make it a lot quicker and easier to develop and test proposed word-sense disambiguation algorithms. Bar-Hillel, of course, was not very supportive of empirical methods which were falling out of favor at the time. He criticized a statistical/ empirical approach like the one that we have been advocating on the grounds that it didn't work very well. At the time, of course, it wasn't practical to train the system on tens of millions of words of parallel text.

> ‘‘Let me state rather dogmatically that there exists at this moment no method of reducing the polysemy of the, say, twenty words of an average Russian sentence in a scientific article below a remainder of, I would estimate, at least five or six words with multiple English renderings, which would not seriously endanger the quality of the machine output. Many tend to believe that by reducing the number of initially possible renderings of a twenty word Russian sentence from a few tens of thousands (which is the approximate number resulting from the assumption that each of the twenty Russian words has two renderings on the average, while seven or eight of them have only one rendering) to some eighty (which would be the number of renderings on the assumption that sixteen words are uniquely rendered and four have three renderings apiece, forgetting now about all the other aspects such as change of word order, etc.) the main bulk of this kind of work has been achieved, the remainder requiring only some slight additional effort.’’ (Bar-Hillel, 1960, p. 163)

Bar-Hillel’s argument is basically an empirical one. Using the numbers mentioned in the quote, it appears that the statistical approach could only disambiguate about 75% of the words in a sentence (e.g., 15 out of 20). This level of performance is really not very good, and would still leave about eighty renderings, as Bar-Hillel points out. However, now that we seem to be achieving 90-95%, the eighty renderings would be reduced to about six renderings (assuming that 18 of the 20 words would be disambiguated and that the remaining two words would have three renderings apiece). At this point, the disambiguation routine would probably be good enough to to facilitate further lexicographic research.

In particular, some of the lexicographers that we have worked with have expressed the desire to sort concordances on word senses, as an alternative to the more usual option of sorting concordance lines alphabetically using the context to the left or the right of the probe. We suspect that these lexicographers would be very happy if 90% of the concordances ended up in the right place. Admittedly even 90% disambiguation is not good enough for all applications. For example, it probably isn’t good enough for FAQHT (fully-automatic high-quality translation) since a large number of sentences would still have more than one rendering.

The debate over statistical methods for sense-disambiguation is similar to the debate over statistical methods for part-of-speech assignment. There was a time when most researchers shyed away from statistical taggers such as (Church, 1988) because they have obvious theoretical limitations. Nevertheless, these methods have since become quite popular since it has become clear that these statistical methods work better than anything else we currently have at our disposal, and they work well enough to meet the needs of many short-term applications. So too, we might suggest that statistical sense-disambiguation algorithms are likely to out-perform most of the currently available alternatives, and they are likely to meet the needs of many short-term applications, even though they may have obvious theoretical limitations.

6. Concordances Based on Word Correspondences

Most currently available bilingual concordance programs (e.g., Table 3, ISSCO’s BCP program mentioned in footnote 1 of (Warwick *et al.*, 1990) and a similar program mentioned on page 20 of (Klavans and Tzoukermann, 1990b)) are based on aligned sentences. These sentence-based programs require the user to supply the program with both an English and a French word (e.g., *drug* and *médicaments*). The problem is that it is often not very easy to for the user to think of all the possible translations for the probe (e.g., *drug*).

In fact, it is possible for the computer to help identify many of these translations. It is generally easier for the user to look at a list of candidate translations and agree that the program’s suggestion is reasonable. Table 5 below shows the ten most frequent ‘‘translations’’ for *drug* and *sentence*, discovered by a self-organizing method which will be discussed in the next section. The symbol ‘‘*NULL*’’ is used when the program was unable to find a corresponding word in the French text.[2]

2. Accents have been removed from the words in this table.

Table 5: Top 10 Choices

	drug		sentence
1023	medicaments	217	peine
297	pharmaceutiques	105	phrase
285	drogues	92	sentence
262	medicament	27	*NULL*
145	drogue	14	condamnation
118	*NULL*	10	condamne
96	pharmaceutique	9	peines
84	assurance-medicaments	9	condamnees
61	toxicomanie	7	emprisonnement
43	stupefiants	6	detenu

This table shows how a word-based concordance tool can be more useful than a sentence-based tool. Suppose, for example, that one was trying to check the precision and recall of a particular bilingual dictionary entry. A sentence-based concordance program would be quite useful for the first part, checking precision (making sure that the statements in the dictionary are generally supported by corpus evidence), but it would be less useful for checking recall (making sure that most of the corpus evidence is covered by statements in the dictionary). A sentence-based bilingual concordance program is not very good at showing what isn't already covered. One could ask for all of the concordance lines matching the English word *drug* and where the French text does not contain one of the more obvious translations, e.g., *drogue*, *médicament*. But the program will not be able to line up the English concordance line with the French concordance line, since , by construction, there is no word in the French text to line up under the English probe (*drug*).

In short, there are two limitations with a sentence-based concordance program. First, the tool requires the user to provide a list of candidate translations of the probe, and secondly, the tool is not very good for looking at the residue, the instances of the probe that do not have a clear correspondence in the translation.

We have found that we can address these concerns by making use of a program that identifies many of the word correspondences and uses them to line up the text. Even when the program doesn't have a correspondence for the probe, it can often do a fairly good job by lining up many of the words in the context as we will see. In this way, it is possible for the concordance program to relieve the user of the responsibility of providing the translations for the probe. Our word-based concordance program takes a probe such as *drug*, prints out the English concordance lines just as in the monolingual case, and then lines them up with the corresponding French text. This program depends crucially on the availability of an index identifying word correspondences between the English and French portions of the corpus. For each token in the English text, there is a pointer to the corresponding word in the French file (or *NULL*, if there isn't a corresponding word in the other text).

Consider, for example, the English sentence, *The drug was simply banned* and its French translation, *Ce dernier a éte simplement interdit*. Note that the French translation does not explicitly mention drugs; *ce dernier* is probably referring to a previous mention of drugs in the discourse. In fact, this kind of discourse dependency is fairly common in the Hansard data.

Even when *drug* does not have a good correspondence in the translation, the word-based concordance program can still do a fairly good job of lining up the text and its translation by making use of nearby words in the concordance line. In this case, the correspondence program is able to find two matches:

The drug was simply banned
simplement interdit

The concordance program can use the correspondence between *simply* and *simplement* to decide how to put the English and French texts into registration.

The drug was | simply banned.
Ce dernier a ete | simplement interdit.

In general, there will usually be several corresponding words in the 100 character window. Each one can be used to predict how to align the two texts. The word-based concordance program averages these predictions together to produce an alignment which is generally close enough so that a lexicographer can look at the paired lines and generally figure out fairly quickly how the probe corresponds to the translation.

7. Identifying Word-Correspondences

7.1 Problem Statement

The program is intended to identify which words in the English text correspond to which words in the French text. For example, given the input sentences:

English: We took the initiative in assessing and amending current legislation and policies to ensure that they reflect a broad interpretation of the charter.

French: Nous avons pris l'initiative d'évaluer et de modifier des lois et des politiques en vigueur afin qu'elles correspondent à une interprétation généreuse de la charte.

The program, described in (Gale and Church, 1991b), would produce the following alignment:

We took the initiative in assessing and amending current
pris *initiative* *evaluer* *modifier*

legislation and policies to ensure that they reflect
lois *politiques* *afin* *correspondent*

a broad interpretation of the charter
genereuse *interpretation* *charte*

In this example, 11 out of the 23 (48%) English words were matched with a French word, and 12 of the English words were left unmatched. This program has been run on the 18 million word sample of Canadian Hansards mentioned above.

We wish to distinguish the terms *alignment* and *correspondence*. The term *alignment* will be used when order constraints must be preserved and the term *correspondence* will be used when order constraints need not be preserved and crossing dependencies are permitted. We refer to the matching problem at the word level as a correspondence problem because it is important to model crossing dependencies (e.g., *sales volume* and *volume des ventes*). In contrast, we refer to the matching problem at the sentence level as an alignment problem because we believe that it is not necessary to model crossing dependencies at the sentence level as they are quite rare and can be ignored (at least for now).

8. Contingency Tables

Suppose that we suspected that *house* and *chambre* were translations of one another, and we wanted to confirm this hypothesis. It turns out that *house* and *chambre* are, in fact, highly associated in this rather unusual genre. In most other genres, *maison* would almost certainly be the preferred translation. Table 6 below shows the ten most frequent translations of *house* and *houses* in the Hansards.[3]

Table 6: Top 10 Choices for ‘‘house’’ and ‘‘houses’’

	house		houses
30,584	chambre	159	maisons
5737	*NULL*	112	chambres
3145	deputes	60	*NULL*
1330	depute	18	maison
281	maison	17	chambre
80	deputee	10	logements
60	house	6	deputes
40	depose	3	habitations
38	deposer	2	logement
34	chambres	2	depot

The choices for the plural form *houses* are probably closer to what one might expect in a balanced corpus of general English. Hopefully we will soon gain access to such a corpus, but in the meantime, we will use the unbalanced Hansards to develop and test the methodology. If we can show that the methods work sufficiently well on these unbalanced texts, then perhaps we can motivate some of you to help us obtain access to more appropriate samples of text.

The *house*/ *chambre* example is a good one for illustrating our use of contingency tables because the pair is so common, at least in our Hansard data. How can we confirm the hypothesis that *house* and *chambre* are translations of one another? Table 7 gives a two-by-two contingency table for the two words. Cell *a* (upper-left) counts the number of sentences (aligned regions) that contain both *house* and *chambre*. Cell *b* (upper-right) counts the number of regions that contain *house* but not *chambre*. Cells *c* and *d* fill out the pattern in the obvious way.

The table can be computed from $freq(house)$, $freq(chambre)$ and $freq(house, chambre)$, the number of aligned regions that contain one or both these words, and from $N = 897{,}077$, the total number of regions.[4]

$$
\begin{aligned}
a &= freq(house, chambre) \\
b &= freq(house) - freq(house, chambre) \\
c &= freq(chambre) - freq(house, chambre) \\
d &= N - a - b - c
\end{aligned}
$$

Table 7: A Contingency Table

	chambre	
house	31,950	12,004
	4,793	848,330

We can now measure the association between *house* and *chambre* by making use of any one of a number of association measures such as mutual information. ϕ^2, a χ^2-like statistic, seems to be a particularly good choice because it makes good use of the off-diagonal cells *b* and *c*.

3. Accents have been removed from the words in this table.

4. The value of 31,950 in this table is slightly larger than the 30,584 reported in the previous table because of an uninteresting detail in processing. The larger value was calculated on all regions, whereas the smaller number was calculated on a subset of the regions with reasonably small distance scores.

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

ϕ^2 is bounded between 0 and 1. In this case, ϕ^2 is 0.62, a relatively high value, indicating the two words are strongly associated, and that they may be translations of one another. One can make this argument more rigorous by measuring the confidence that ϕ^2 is different from chance. In this case, the variance of ϕ^2 is estimated to be 2.5×10^{-5} (see section 7.4), and hence

$$t = \frac{\phi^2}{\sqrt{var(\phi^2)}} = \frac{0.62}{\sqrt{2.5 \times 10^{-5}}} = 123$$

With such a large *t*, we can very confidently reject the null hypothesis and assume that there is very likely to be an association between *house* and *chambre*.

8.1 A Near Miss

One might contrast *house*/*chambre* with a near miss such as *house*/*communes* (see Table 8). For the sake of discussion, let us assume that *House of Commons* and *Chambre des Communes* should come out as

House of Commons
Chambre des Communes

and not as

*House of Commons
Communes des Chambre

This example is intended to illustrate that association statistics often have difficulty distinguishing translations of the probe from translations of collocates of the probe. Most association statistics depend on the noise being uncorrelated with the signal. Unfortunately, collocates are highly correlated with the signal, so they can be very distracting.

Table 8 below gives the two-by-two contingency table for *house* and *communes*. We might hope that these two words would have a small ϕ^2 since they are not translations of each other.

Table 8: A Near Miss

	communes	
house	4,974	38,980
	441	852,682

Unfortunately, this pair is also highly significant ($t = 31$), because *house* and *communes* do, in fact, co-occur together significantly more than chance. How do we know that *house* is more associated with *chambre* than with *communes*? Note that mutual information does not distinguish these two pairs. Recall the mutual information $I(x;y)$ is computed by

$$\log_2 \frac{Pr(x,y)}{Pr(x)Pr(y)}$$

where $Pr(x,y) = a/N$, $Pr(x) = (a + b)/N$, $Pr(y) = (a + c)/N$ and $N = a + b + c + d$. If we plug these numbers into the formulas, we find that *house* and *chambre* actually have a lower mutual information value than *house* and *communes*.

$$I(house;chambre) = 4.1$$

$$I(house;communes) = 4.2$$

Mutual information picks up the fact that there are strong associations in both cases. Unfortunately, it is not very good at deciding which association is stronger. In addition, it does not make very good use of the off-diagonal cells *b* and *c*, which are often better estimated than cell *a* since the counts in *b* and *c* are often larger than those in *a*.

In this case, the crucial difference is that cell *b* is much smaller in Table 7 than in Table 8. ϕ^2 picks up this difference; Table 8 has a ϕ^2 of 0.098, much less than Table 7's ϕ^2 of 0.62. The difference is highly significant ($t = 88$). Thus, we can very confidently say that *house* is more associated with *chambre* than with *communes*.

8.2 Calculation of Variances

The estimate of $var(\phi^2)$ is very important to this argument. We use the following reasoning:

$$var(\phi^2) \approx \left[\frac{\partial \phi^2}{\partial a}\right]^2 var(a) + \left[\frac{\partial \phi^2}{\partial b}\right]^2 var(b) + \left[\frac{\partial \phi^2}{\partial c}\right]^2 var(c) + \left[\frac{\partial \phi^2}{\partial d}\right]^2 var(d)$$

where $var(a) \approx a$, $var(b) \approx b$, $var(c) \approx c$ and $var(d) \approx a + b + c$. A direct calculation of this is valid when ϕ^2 is small. That is, $var_{small}(\phi^2) \approx$

$$4\,\phi^2\,\frac{a^2 var(d) + d^2 a + b^2 c + c^2 b}{(a+b)(c+d)(a+c)(b+d)} +$$

$$\phi^4\left(\frac{1}{a+b} + \frac{c + var(d)}{(c+d)^2} + \frac{1}{a+c} + \frac{b + var(d)}{(b+d)^2}\right)$$

As ϕ^2 approaches 1, var_{small} decreases to 0, which makes this equation unsuitable as an estimate of the variance. We calculate a variance for this case by assuming that $bc << ad$, which implies that $\phi^2 \approx 1 - (b + c)/a$. With this assumption, we obtain

$$var_{large}(\phi^2) \approx a^{-2}\,(b + c)\,\left(1 + \frac{b + c}{a}\right)$$

We do not have an exact relation to specify when ϕ^2 is large and when it is small. Rather, we observe that each estimate produces a value that is small in its domain, so we estimate the variance of ϕ^2 by the minimum of the two cases:

$$var(\phi^2) \approx \min\,(var_{small}, var_{large})$$

8.3 Selecting Pairs

We have now seen how we could decide that *house* and *chambre* are more associated than *house* and *communes*. But why did we decide to look at these pairs of words and not some others? We probably can't

afford to look at all possible English and French words, unless we limit the vocabulary sizes down to something like the 9000 word limit in Brown *et al.* (1990). And even then, there would be 81 million pairs to consider. If the training corpus is not too large (e.g., 50,000 regions), then it is possible to consider all pairs of words that actually co-occur in at least one region (i.e., $a \neq 0$). Unfortunately, with a training corpus of $N = 890{,}000$ regions, we have found that there are too many such pairs and it becomes necessary to be more selective (heuristic).

Fortunately, it is fairly easy to identify many candidate pairs after looking at a relatively small subset of the training corpus. The *house*/ *chambre* pair, for example, is so frequent that one might suspect that there is a relationship after seeing just a few examples. After looking at just a handful of examples, we might already suspect that *house* and *chambre* deserve further investigation.

We have had fairly good success with a progressive deepening strategy. That is, select a small set of regions (e.g., 10,000) and use all of the training material to compute ϕ^2 for pairs of words that appear in any of these 10,000 regions. Select the best pairs. That is, take a pair (x, y) if it has a ϕ^2 significantly better than any other pair of the form (x, z) or (z, y). This procedure would take *house*/ *chambre* but not *house*/ *communes*. Repeat this operation with larger and larger samples of the training corpus. On each iteration, remove pairs of words from the training corpus that have already been selected so that other alternatives can be identified.

8.4 Returning to the Sentence Context

It is now time to try to put these pairs back into their sentence context. Consider the example mentioned previously:

> *English*: We took the initiative in assessing and amending current legislation and policies to ensure that they reflect a broad interpretation of the charter.
>
> *French*: Nous avons pris l'initiative d'évaluer et de modifier des lois et des politiques en vigueur afin qu'elles correspondent à une interprétation généreuse de la charte.

The matching procedure attempts to match English and French words using pairs by the procedure described in the previous section. (Many function words are put on a stop list in order to avoid trouble.) When there are multiple possibilities, the procedure uses a slope condition to select the best pair. A few slopes are shown below:

We	took	the	initiative	in	assessing	and	amending	current
pris			*initiative*		*evaluer*		*modifier*	
3			2		2		3	

There is a slope of 3 under *amending*, for example, because there are three French words (*et de modifier*) between the translation of *initiative*, the previously matched English word, and the translation of the current English word *amending*. The other slopes in the example above are computed by the same reasoning.

We empirically determined a probability for each slope, and use dynamic programming to find the sequence of matches that minimize the product of the slope probabilities.

9. Extensions for Complicated Words

We have extended the concordance tool in a number of ways that are especially helpful when analyzing complicated words like *take*. As is well-known, *take* is often used in complex idiomatic expressions, and consequently, there are many uses of *take* that should not be translated with *prendre*. In fact, most uses of *take* are not translated with *prendre* (or one of its morphological variants). The word-based bilingual concordances show this fairly clearly. Only one of the first ten uses of *take* is translated with *prendre*. The striking absence of *prendre* is consistent with the observation in the Cobuild dictionary (Sinclair *et al.*,

1987, p. 1488) that ‘‘[t]he most frequent use of **take** is in expressions where it does not have a very distinct meaning of its own, but where most of the meaning is in... the direct object.’’

It is natural to hypothesize that *take* would translate differently in some of these complex expressions. The tool has been generalized to allow us to check this kind of hypothesis. The probe is no longer required to be just a word, but can now be specified with an expression such as

(meet take advantage 1 2)

which will find all instances of *take* followed by *advantage*. There happen to be 412 of these in our Hansard data. Not surprisingly, these are almost always translated as *profiter*, and hardly ever as *prendre*. Moreover, the word *advantage* is rarely translated explicitly. In fact, only 28 of the 412 instances mention the French word *avantage* or some variant thereof.

The tool also allows joins. For example, the expression

(meet take (union advantage opportunity) 3 5)

would find all instances of *take* where *advantage* or *opportunity* appear within 3 words to the left or 5 words to the right.

Since many of these patterns would return many more concordance lines than a lexicographer can deal with easily, we have added the *sample* operator. Thus, for example,

(sample (meet take advantage 1 2) 10)

would return ten randomly selected examples of *take advantage*. This is a very useful option for getting a quick sense of the lay of the land.

In some cases, the user might want to specify a particular translation. Thus, for example,

(trans (sample risk 500) jeu)

will generate a sample of 500 instances of *risk* and return those that happen to be translated as *jeu*. It happens that *at risk* is often translated as *en jeu* (more literally, *in play*), a possibility that is not mentioned in the Collins Robert English-French French-English dictionary (Atkins *et al.*, 1987). We recommend using the *sample* operator inside of the *trans* operator to save time. The *trans* operation is somewhat slow because it generally needs to go to the disk once for each instance of its first argument.

The *nottrans* operation is provided to eliminate unwanted translations. For example,

(nottrans (sample risk 50) (union risque danger))

would select examples of *risk* that are not translated as *risque* or *danger*. Although this function is relatively easy to implement, and many people have requested this capability, we have generally not found it to be very useful because it is very hard to think of all the crap that you don’t want to look at. We recommend, instead, that one sort the output on the French translation of the probe. This has the effect of lumping most of the crap together in one place, so that the user can quickly skim over it.

The *notrans* function is provided to find all those instances of the probe which correspond to ‘‘*NULL*’’. We have found this useful in debugging the correspondence program. Lexicographers may or may not find this useful; it generally returns a mixed bag of odd-ball cases that demonstrate just how flawed the word-for-word assumption really is.

10. Risk and Ignore

For a complex ‘‘function’’ word like *take*, it is clear that the word-based concordance tool may be more useful than the sentence-based approach since the lexicographer is hard-pressed to come up with a good list of candidate translations. Even for a moderately transparent word like *risk* or *ignore*, the word-based concordance tool shows some very interesting patterns that might not have been obvious. For example, as mentioned above, there is fairly clear evidence that *at risk* can be translated as *en jeu*, a possibility that is not mentioned in (Atkins *et al.*, 1987). Table 9 shows some common ‘‘translations’’ of *risk* and *ignore*.[5]

Table 9: Common Translation of ‘‘risk’’ and ‘‘ignore’’

	risk		ignore
363	risque	108	compte
188	risques	90	*NULL*
135	danger	47	ignorer
98	*NULL*	43	fi
55	peril	11	fermer
45	menaces	10	negliger
38	risquent		
25	risquer		
24	jeu		
21	menacees		
16	menace		
16	menacee		
15	compromettre		
13	risquons		
11	compromis		
11	securite		
10	dangers		

Admittedly, the program is not perfect. For example, the 11 cases of *sécurité* illustrate that the program is still having some trouble with collocations. In addition, the 108 cases of *compte* illustrate that the word-for-word hypothesis is flawed. Many of these are really *ne pas tenir compte de*. Nevertheless, it shouldn’t be too hard for a lexicographer to look at this table, or at the bilingual concordance lines sorted by translation, and quickly skim past the junk.

After looking at a table such as Table 9 above, one might want to revise the entry for *risk* in a bilingual dictionary. In addition to a few possible omissions such as *compromettre* and *en jeu*, one might also want to reallocate the allotment of space to better reflect the frequency of use. The word *menace*, for example, seems to be a fairly common translation of *risk*, as least in this genre, as Table 9 shows; perhap it should be given more space than the very brief mention that it receives in (Atkins *et al.*, 1987). In general, perhaps frequency arguments should play a greater role in bilingual dictionaries, for many of the same reasons that led to their use in leaners’ dictionaries such as Cobuild (Sinclair *et al.*, 1987).

The English word *ignore* is an interesting example because it doesn’t have a very good French translation. Evidently, the French word *ignorer* is a false cognate, though it is used fairly often in the Hansards as a translation of *ignore*, nevertheless. A number of French speakers have asked us about *fi*. Apparently, the expression *faire fi de* is fairly common in Quebec, at least in this genre, although it is sufficiently rare in other genres and in other dialects of French that many well-educated French speakers have never heard of it.

5. Accents have been removed from the words in this table.

Of course, the Hansards are not a balanced sample of general French.[6] Ideally, one would prefer to work with some text that covered a wider range of genre's and dialects, etc. Nevertheless, we have found the resource to be much more useful than one might have thought. It would be a serious mistake to dismiss the great opportunities that parallel corpora have to offer just because it is currently hard to obtain a balanced sample of parallel text. We have found that text is generally becoming easier and easier to obtain. If the research community continues to develop more and better ways to make use of these texts, then the resources will be made available to collect whatever text we need.

11. Conclusions

We have described two types of bilingual concordance tools, sentence-based and word-based. The simpler sentence-based method makes use of pre-computed index of sentence alignments, indicating which sentences in the one language correspond to which sentences in the other language. The sentence-based concordance program takes two words as input: (1) a probe in one language (e.g., the English word *risk*) and (2) a translation of the probe in the other language (e.g., *risque*). The program searches sentences that contain the probe, and whose translation also contains the translation of the probe. The most difficult part of implementing this tool is the construction of the index of sentence alignments, which was discussed in section 2.

The word-based concordance method makes use of a pre-computed index indicating which words in the one language correspond to which words in the other language. The word-based concordance program takes just one word as input (e.g., *risk*), and uses the index of word correspondences to align concordance lines with their translations. In this case, the most difficult implementation issue is the construction of the index of word correspondences which was discussed in section 7.

Kenneth Church

William Gale

6. Sports metaphors are a particularly rich source of regional language. In particular, terms from hockey seem to be particularly specific to Canadian dialects of both English and French. For example, the expression *ragging the puck* (*faire trainer la rondelle*) refers to what hockey players do to kill a penalty. Apparently, it is also used quite often in the Hansards, especially when one member of parliament wants to accuse another member of waisting time with great skill. There are quite a number of other references to *puck* (e.g., *losing the puck, putting the puck in the net*), most of which seem to have a special meaning in Canada. The word *puck* apparently isn't important enough in Europe to be listed in (Atkins *et al.*, 1987); the French equivalent, *rondelle*, is listed, but only in the sense of a round piece of food (e.g., a slice of sausage). These hockey metaphors probably have somewhat the same status as our baseball metaphors (e.g., *drop the ball, out in left field*), which also don't travel very well.

References

Atkins, B., *et al.* (1987), *Collins Robert French-English English-French Dictionary, Second Edition*, Collins.

Bar-Hillel, Y. (1960) ‘‘The Present Status of Automatic Translation of Languages,’’ in F. Alt (ed.), *Advances in Computers*, Vol. 1, Academic Press, New York.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, J. Lafferty, R. Mercer, and P. Roossin (1990) ‘‘A Statistical Approach to Machine Translation,’’ *Computational Linguistics*, v 16, pp 79-85.

Brown, P, J. Lai, R. Mercer (1991a) ‘‘Aligning Sentences in Parallel Corpora,’’ ACL.

Brown, P., Dela Pietra, S., Dela Pietra, V., Mercer, R. (1991b), ‘‘Word Sense Disambiguation using Statistical Methods,’’ ACL.

Catizone, R., G. Russell, and S. Warwick (to appear) ‘‘Deriving Translation Data from Bilingual Texts,’’ in Zernik (ed), *Lexical Acquisition: Using on-line Resources to Build a Lexicon*, Lawrence Erlbaum.

Chomsky, N. (1957) *Syntactic Structures*, The Hague: Mouton & Co.

Church, K., (1988) ‘‘A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text,’’ Second Conference on Applied Natural Language Processing, Austin, Texas.

Gale, W. and K. Church (1991a) ‘‘A Program for Aligning Sentences in Bilingual Corpora,’’ ACL.

Gale, W. and K. Church (1991b) ‘‘Identifying Word Correspondences in Parallel Texts,’’ Darpa SLS Workshop.

Kay, M. and M. Röscheisen (1988) ‘‘Text-Translation Alignment,’’ unpublished ms., Xerox Palo Alto Research Center.

Klavans, J., and E. Tzoukermann (1990a) ‘‘The BICORD System,’’ *COLING-90*, pp 174-179.

Klavans, J., and E. Tzoukermann (1990b) ‘‘Linking Bilingual Corpora and Machine Readable Dictionaries with the BICORD System,’’ Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, pp. 19-30.

Salton, G. (1989) *Automatic Text Processing*, Addison-Wesley Publishing Co.

Sinclair, J., Hanks, P., Fox, G., Moon, R., Stock, P. et al. (eds.) (1987) *Collins Cobuild English Language Dictionary*, Collins, London and Glasgow.

Warwick, S. and G. Russell (1990) ‘‘Bilingual Concordancing and Bilingual Lexicography,’’ Euralex 1990.

Warwick, S., J. Hajic and G. Russell (1990) ‘‘Searching on Tagged Corpora: Linguistically Motivated Concordance Analysis,’’ Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research, pp. 10-18.