A PROGRAM FOR ALIGNING SENTENCES IN BILINGUAL CORPORA

William A. Gale
Kenneth W. Church

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ, 07974

ABSTRACT

Researchers in both machine translation (e.g., Brown *et al.*, 1990) and bilingual lexicography (e.g., Klavans and Tzoukermann, 1990) have recently become interested in studying parallel texts, texts such as the Canadian Hansards (parliamentary proceedings) which are available in multiple languages (French and English). This paper describes a method for aligning sentences in these parallel texts, based on a simple statistical model of character lengths. The method was developed and tested on a small trilingual sample of Swiss economic reports. A much larger sample of 90 million words of Canadian Hansards has been aligned and donated to the ACL/DCI.

1. Introduction

Researchers in both machine translation (e.g., Brown et al, 1990) and bilingual lexicography (e.g., Klavans and Tzoukermann, 1990) have recently become interested in studying bilingual corpora, bodies of text such as the Canadian Hansards (parliamentary debates) which are available in multiple languages (such as French and English). The sentence alignment task is to identify correspondences between sentences in one language and sentences in the other language. This task is a first step toward the more ambitious task finding correspondances among words.[1]

The input is a pair of texts such as Table 1.

Table 1:
Input to Alignment Program

English

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. The higher turnover was largely due to an increase in the sales volume. Employment and investment levels also climbed. 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.

French

Quant aux eaux minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment. La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes. L'emploi et les investissements ont également augmenté. La nouvelle ordonnance fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

The output identifies the alignment between sentences. Most English sentences match exactly one French sentence, but it is possible for an English sentence to match two or more French sentences. The first two English sentences (below) illustrate a particularly hard case where two English sentences align to two French sentences. No smaller alignments are possible because the clause "... sales ... were higher..." in

1. In statistics, string matching problems are divided into two classes: *alignment* problems and *correspondance* problems. Crossing dependencies are possible in the latter, but not in the former.

the first English sentence corresponds to (part of) the second French sentence. The next two alignments below illustrate the more typical case where one English sentence aligns with exactly one French sentence. The final alignment matches two English sentences to a single French sentence. These alignments agreed with the results produced by a human judge.

Table 2:
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 minérales et aux limonades, elles rencontrent toujours plus d'adeptes. En effet, notre sondage fait ressortir des ventes nettement supérieures à celles de 1987, pour les boissons à base de cola notamment.
The higher turnover was largely due to an increase in the sales volume.
La progression des chiffres d'affaires résulte en grande partie de l'accroissement du volume des ventes.
Employment and investment levels also climbed.
L'emploi et les investissements ont également 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 fédérale sur les denrées alimentaires concernant entre autres les eaux minérales, entrée en vigueur le 1er avril 1988 après une période transitoire de deux ans, exige surtout une plus grande constance dans la qualité et une garantie de la pureté.

Aligning sentences is just a first step toward constructing a probabilistic dictionary (Table 3) for use in aligning words in machine translation (Brown et al., 1990), or for constructing a bilingual concordance (Table 4) for use in lexicography (Klavans and Tzoukermann, 1990).

Table 3:
An Entry in a Probabilistic Dictionary
(from Brown et al., 1990)

English	French	Prob(French | English)
the	le	0.610
the	la	0.178
the	l'	0.083
the	les	0.023
the	ce	0.013
the	il	0.012
the	de	0.009
the	à	0.007
the	que	0.007

Table 4: A Bilingual Concordance

bank/banque ("money" sense)	
and the governor of the	bank of canada have frequently
et le gouverneur de la	banque du canada ont fréquemm
800 per cent in one week through	bank action . SENT there
% en une semaine à cause d' une	banque . SENT voilà
bank/banc ("place" sense)	
such was the case in the georges	bank issue which was settled betw
ats-unis et le canada à propos du	banc de george .
he said the nose and tail of the	bank were surrendered by
cédé les extrémités du	banc . SENT en fait

Although there has been some previous work on the sentence alignment, e.g., (Brown, Lai, and Mercer, 1991), (Kay and Röscheisen, 1988), (Catizone et al., to appear), the alignment task remains a significant obstacle preventing many potential users from reaping many of the benefits of bilingual corpora, because the proposed solutions are often unavailable, unreliable, and/or computationally prohibitive.

The *align* program is based on a very simple statistical model of character lengths. The model makes use of the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences. A probabilistic score is assigned to each pair of proposed sentence pairs, based on the ratio of lengths of the two sentences (in characters) and the variance of this ratio. This probabilistic score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences.

It is remarkable that such a simple approach can work as well as it does. An evaluation was performed based on a trilingual corpus of 15 economic reports issued by the Union Bank of Switzerland (UBS) in English, French and German (N = 14,680 words, 725 sentences, and 188 paragraphs in English and corresponding numbers in the other two languages). The method correctly aligned all but 4% of the sentences. Moreover, it is 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 is reduced from 4% to 0.7%. There were roughly the same number of errors in each of the English-French and English-German alignments, suggesting that the method may be fairly language independent. We believe that the error rate is considerably lower in the Canadian Hansards because the translations are more literal.

2. A Dynamic Programming Framework

Now, let us consider how sentences can be aligned within a paragraph. The program makes use of the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences.[2] A probabilistic score is assigned to each proposed pair of sentences, based on the ratio of lengths of the two sentences (in characters) and the variance of this ratio. This probabilistic score is used in a dynamic programming framework in order to find the maximum likelihood alignment of sentences. We were led to this approach after noting that the lengths (in characters) of English and German paragraphs are highly correlated (.991), as illustrated in the following figure.

Paragraph Lengths are Highly Correlated

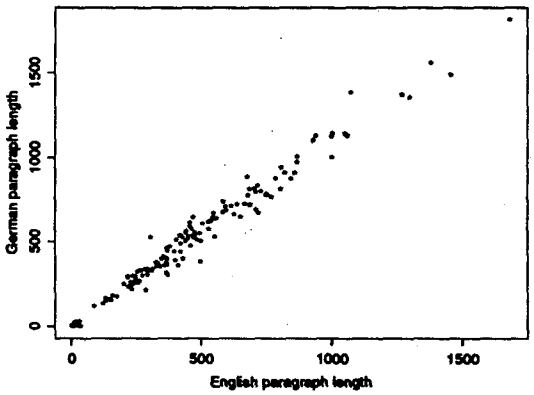

Figure 1. The horizontal axis shows the length of English paragraphs, while the vertical scale shows the lengths of the corresponding German paragraphs. Note that the correlation is quite large (.991).

Dynamic programming is often used to align two sequences of symbols in a variety of settings, such as genetic code sequences from different species, speech sequences from different speakers, gas chromatograph sequences from different compounds, and geologic sequences from different locations (Sankoff and Kruskal, 1983). We could expect these matching techniques to be useful, as long as the order of the sentences does not differ too radically between the two languages. Details of the alignment techniques differ considerably from one application to another, but all use a distance measure to compare two individual elements within the sequences, and a dynamic programming algorithm to minimize the total distances between aligned elements within two sequences. We have found that the sentence alignment problem fits fairly well into this framework.

2. We will have little to say about how sentence boundaries are identified. Identifying sentence boundaries is not always as easy as it might appear for reasons described in Liberman and Church (to appear). It would be much easier if periods were always used to mark sentence boundaries, but unfortunately, many periods have other purposes. In the Brown Corpus, for example, only 90% of the periods are used to mark sentence boundaries; the remaining 10% appear in numerical expressions, abbreviations and so forth. In the Wall Street Journal, there is even more discussion of dollar amounts and percentages, as well as more use of abbreviated titles such as *Mr.*; consequently, only 53% of the periods in the the Wall Street Journal are used to identify sentence boundaries.

   For the UBS data, a simple set of heuristics were used to identify sentences boundaries. The dataset was sufficiently small that it was possible to correct the remaining mistakes by hand. For a larger dataset, such as the Canadian Hansards, it was not possible to check the results by hand. We used the same procedure which is used in (Church, 1988). This procedure was developed by Kathryn Baker (private communication).

3. The Distance Measure

It is convenient for the distance measure to be based on a probabilistic model so that information can be combined in a consistent way. Our distance measure is an estimate of $-log\ Prob(match|\delta)$, where δ depends on l_1 and l_2, the lengths of the two portions of text under consideration. The log is introduced here so that adding distances will produce desirable results.

This 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 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. We define δ to be $(l_2 - l_1\ c)/\sqrt{l_1\ s^2}$ so that it has a normal distribution with mean zero and variance one (at least when the two portions of text under consideration actually do happen to be translations of one another).

The parameters c and s^2 are determined empirically from the UBS data. We could estimate c by counting the number of characters in German paragraphs then dividing by the number of characters in corresponding English paragraphs. We obtain $81105/73481 \approx 1.1$. The same calculation on French and English paragraphs yields $c \approx 72302/68450 \approx 1.06$ as the expected number of French characters per English characters. As will be explained later, performance does not seem to very sensitive to these precise language dependent quantities, and therefore we simply assume $c \approx 1$, which simplifies the program considerably.

The model assumes that s^2 is proportional to length. The constant of proportionality is determined by the slope of a robust regression. The result for English-German is $s^2 = 7.3$, and for English-French is $s^2 = 5.6$. Again, we have found that the difference in the two slopes is not too important. Therefore, we can combine the data across languages, and adopt the simpler language independent estimate $s^2 \approx 6.8$, which is what is actually used in the program.

We now appeal to Bayes Theorem to estimate $Prob(match|\delta)$ as a constant times $Prob(\delta|match)\ Prob(match)$. The constant can be ignored since it will be the same for all proposed matches. The conditional probability $Prob(\delta|match)$ can be estimated by

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

where $Prob(|\delta|)$ is the probability that a random variable, z, with a standardized (mean zero, variance one) normal distribution, has magnitude at least as large as $|\delta|$

The program computes δ directly from the lengths of the two portions of text, l_1 and l_2, and the two parameters, c and s^2. That is, $\delta = (l_2 - l_1\ c)/\sqrt{l_1\ s^2}$. Then, $Prob(|\delta|)$ is computed by integrating a standard normal distribution (with mean zero and variance 1). Many statistics textbooks include a table for computing this.

The prior probability of a match, $Prob(match)$, is fit with the values in Table 5 (below), which were determined from the UBS data. We have found that a sentence in one language normally matches exactly one sentence in the other language (1-1), three additional possibilities are also considered: 1-0 (including 0-1), 2-1 (including 1-2), and 2-2. Table 5 shows all four possibilities.

Table 5: Prob(match)

Category	Frequency	Prob(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
	1312	1.00

This completes the discussion of the distance measure. $Prob(match|\delta)$ is computed as an (irrelevant) constant times $Prob(\delta|match)\ Prob(match)$. $Prob(match)$ is computed using the values in Table 5. $Prob(\delta|match)$ is computed by assuming that $Prob(\delta|match) = 2\ (1 - Prob(|\delta|))$, where $Prob(|\delta|)$ has a standard normal distribution. We first calculate δ as $(l_2 - l_1\ c)/\sqrt{l_1\ s^2}$ and then $Prob(|\delta|)$ is computed by integrating a standard normal distribution.

The distance function *two_side_distance* 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 *two_side_distance*$(x_1, y_1; 0, 0)$ be the cost of substituting x_1 with y_1,
2. *two_side_distance*$(x_1, 0; 0, 0)$ be the cost of deleting x_1,
3. *two_side_distance*$(0, y_1; 0, 0)$ be the cost of insertion of y_1,
4. *two_side_distance*$(x_1, y_1; x_2, 0)$ be the cost of contracting x_1 and x_2 to y_1,
5. *two_side_distance*$(x_1, y_1; 0, y_2)$ be the cost of expanding x_1 to y_1 and y_2, and
6. *two_side_distance*$(x_1, y_1; x_2, y_2)$ be the cost of merging x_1 and x_2 and matching with y_1 and y_2.

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 (*two_side_distance*) described in the previous section, and let $D(i,j)$ be the minimum distance between sentences $s_1, \cdots s_i$ and their translations $t_1, \cdots t_j$, under the maximum likelihood alignment. $D(i,j)$ is computed recursively, where the recurrence minimizes 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 calculated by the following recurrence with the initial condition $D(i, j) = 0$.

$$D(i, j) =$$

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

5. Evaluation

To evaluate *align*, its results were compared with a human alignment. 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 having 230 sentences (out of 118 total paragraphs with 725 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 purposed 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 percent of all paragraphs) contained a relatively large fraction of the program's errors (82 percent). Thus, there does seem 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 just the first two. (The third pair is probably 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%).

Table 6: Complex Matches are More Difficult

category	English-French			English-German			total		
	N	err	%	N	err	%	N	err	%
1-0 or 0-1	8	8	100	5	5	100	13	13	100
1-1	542	14	2.6	625	9	1.4	1167	23	2.0
2-1 or 1-2	59	8	14	58	2	3.4	117	10	9
2-2	9	3	33	6	2	33	15	5	33
3-1 or 1-3	1	1	100	1	1	100	2	2	100
3-2 or 2-3	1	1	100	0	0	0	1	1	100

Table 6 breaks down the errors by category, illustrating that complex matches are more difficult. 1-1 alignments are by far the easiest. The 2-1 alignments, which come next, have four times the error rate for 1-1. The 2-2 alignments are harder still, but a majority of the alignments are found. The 3-1 and 3-2 alignments are not even considered by the algorithm, so naturally all three are counted as errors. The most embarrassing category is 1-0, which was never handled correctly. In addition, when the algorithm assigns a sentence to the 1-0 category, it is also always wrong. Clearly, more work is needed to deal with the 1-0 category. It may be necessary to consider language-specific methods in order to deal adequately with this case.

We observe that the score is a good predictor of performance, and therefore the score can be used 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%. In general, we can trade off the size of the subcorpus and the accuracy by setting a threshold, and rejecting alignments with a score above this threshold. Figure 2 examines this trade-off in more detail.

Extracting a Subcorpus with Lower Error Rate

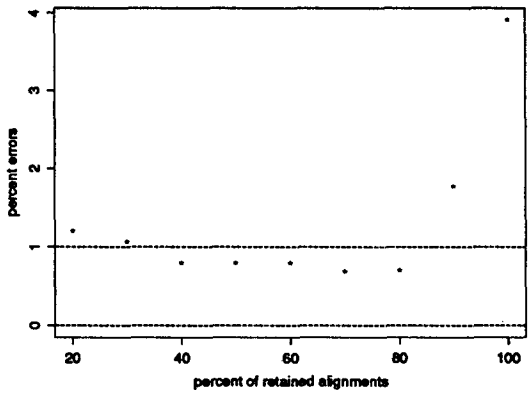

Figure 2. The fact that the score is such a good predictor of performance can be used to extract a large subcorpus which has a much smaller error rate. In general, we can trade-off the size of the subcorpus and the accuracy by setting a threshold, and rejecting alignments with a score above this threshold. The horizontal axis shows the size of the subcorpus, and the vertical axis shows the corresponding error rate. An error rate of about 2/3% can be obtained by selecting a threshold that would retain approximately 80% of the corpus.

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, while the other half of the time is spent in the *align* program itself.

6. Measuring Length in Terms Of Words Rather than Characters

It is interesting to consider what happens if we change our definition of length to count words rather than characters. It might seem that words are a more natural linguistic unit than characters (Brown, Lai and Mercer, 1991). However, we have found that words do not perform nearly as well as characters. In fact, the "words" variation increases the number of errors dramatically (from 36 to 50 for English-French and from 19 to 35 for English-German). The total errors were thereby increased from 55 to 85, or from 4.2% to 6.5%.

We believe that characters are better because there are more of them, and therefore there is less uncertainty. On the average, there are 117 characters per sentence (including white space) and only 17 words per sentence. Recall that we have modeled variance as proportional to sentence length, $V = s^2 \, l$. Using the character data, we found previously that $s^2 \approx 6.5$. The same argument applied to words yields $s^2 \approx 1.9$. For comparison sake, it is useful to consider the ratio of $\sqrt{(V(m))/m}$ (or equivalently, $s/\sqrt{m}$), where m is the mean sentence length. We obtain $\sigma(m)/m$ ratios of 0.22 for characters and 0.33 for words, indicating that characters are less noisy than words, and are therefore more suitable for use in *align*.

7. Conclusions

This paper has proposed a method for aligning sentences in a bilingual corpus, based on a simple probabilistic model, described in Section 3. The model was motivated by the observation that longer regions of text tend to have longer translations, and that shorter regions of text tend to have shorter translations. In particular, we found that the correlation between the length of a paragraph in characters and the length of its translation was extremely high (0.991). This high correlation suggests that length might be a strong clue for sentence alignment.

Although this method is extremely simple, it is also quite accurate. Overall, there was a 4.2% error rate on 1316 alignments, averaged over both English-French and English-German data. In addition, we find that the probability score is a good predictor of accuracy, and consequently, it is possible to select a subset of 80% of the alignments with a much smaller error rate of only 0.7%.

The method is also fairly language-independent. Both English-French and English-German data were processed using the same parameters. If necessary, it is possible to fit the six parameters in

the model with language-specific values, though, thus far, we have not found it necessary (or even helpful) to do so.

We have examined a number of variations. In particular, we found that it is better to use characters rather than words in counting sentence length. Apparently, the performance is better with characters because there is less variability in the ratios of sentence lengths so measured. Using words as units increases the error rate by half, from 4.2% to 6.5%.

In the future, we would hope to extend the method to make use of lexical constraints. However, it is remarkable just how well we can do without such constraints. We might advocate the simple character length alignment procedure as a useful first pass, even to those who advocate the use of lexical constraints. The character length procedure might complement a lexical constraint approach quite well, since it is quick but has some errors while a lexical approach is probably slower, though possibly more accurate. One might go with the character length procedure when the distance scores are small, and back off to a lexical approach as necessary.

ACKNOWLEDGEMENTS

We thank Susanne Wolff and and Evelyne Tzoukermann for their pains in aligning sentences. Susan Warwick provided us with the UBS trilingual corpus and posed the problem addressed here.

REFERENCES

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, and R. Mercer, (1991) "Aligning Sentences in Parallel Corpora," ACL Conference, Berkeley.

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.

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

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

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

Liberman, M., and K. Church, (to appear), "Text Analysis and Word Pronunciation in Text-to-Speech Synthesis," in Furui, S., and Sondhi, M. (eds.), *Advances in Speech Signal Processing*.