Text-Translation Alignment

Martin Kay*
Xerox Palo Alto Research Center
and
Stanford University

Martin Röscheisen†
Xerox Palo Alto Research Center
and
Technical University of Munich

We present an algorithm for aligning texts with their translations that is based only on internal evidence. The relaxation process rests on a notion of which word in one text corresponds to which word in the other text that is essentially based on the similarity of their distributions. It exploits a partial alignment of the word level to induce a maximum likelihood alignment of the sentence level, which is in turn used, in the next iteration, to refine the word level estimate. The algorithm appears to converge to the correct sentence alignment in only a few iterations.

1. The Problem

To *align* a text with a translation of it in another language is, in the terminology of this paper, to show which of its parts are translated by what parts of the second text. The result takes the form of a list of pairs of items—words, sentences, paragraphs, or whatever—from the two texts. A pair $\langle a, b \rangle$ is on the list if a is translated, in whole or in part, by b. If $\langle a, b \rangle$ and $\langle a, c \rangle$ are on the list, it is because a is translated partly by b, and partly by c. We say that the alignment is *partial* if only some of the items of the chosen kind from one or other of the texts are represented in the pairs. Otherwise, it is *complete*.

It is notoriously difficult to align good translations on the basis of words, because it is often difficult to decide just which words in an original are responsible for a given one in a translation and, in any case, some words apparently translate morphological or syntactic phenomena rather than other words. However, it is relatively easy to establish correspondences between such words as proper nouns and technical terms, so that partial alignment on the word level is often possible. On the other hand, it is also easy to align texts and translations on the sentence or paragraph levels, for there is rarely much doubt as to which sentences in a translation contain the material contributed by a given one in the original.

The growing interest in the possibility of automatically aligning large texts is attested to by independent work that has been done on it since the first description of our methods was made available (Kay and Röscheisen 1988). In recent years it has been possible for the first time to obtain machine-readable versions of large corpora of text with accompanying translations. The most striking example is the Canadian "Hansard," the transcript of the proceedings of the Canadian parliament. Such bilingual corpora make it possible to undertake statistical, and other kinds of empirical, studies of translation on a scale that was previously unthinkable.

Alignment makes possible approaches to partially, or completely, automatic translation based on a large corpus of previous translations that have been deemed accept-

* Xerox PARC, 3333 Coyote Hill Road, Palo Alto, CA 94306.
† Department of Computer Science, Technical University of Munich, 8000 Munich 40, Germany.

able. Perhaps the best-known example of this approach is to be found in Sato and Nagao (1990). The method proposed there requires a database to be maintained of the syntactic structures of sentences together with the structures of the corresponding translations. This database is searched in the course of making a new translation for examples of previous sentences that are like the current one in ways that are relevant for the method. Another example is the completely automatic, statistical approach to translation taken by the research group at IBM (Brown et al. 1990), which takes a large corpus of text with aligned translations as its point of departure.

It is widely recognized that one of the most important sources of information to which a translator can have access is a large body of previous translations. No dictionary or terminology bank can provide information of comparable value on topical matters of possibly intense though only transitory interest, or on recently coined terms in the target language, or on matters relating to house style. But such a body of data is useful only if, once a relevant example has been found in the source language, the corresponding passage can be quickly located in the translation. This is simple only if the texts have been previously aligned. Clearly, what is true of the translator is equally true of others for whom translations are a source of primary data, such as students of translation, the designers of translations systems, and lexicographers. Alignment would also facilitate the job of checking for consistency in technical and legal texts where consistency constitutes a large part of accuracy.

In this paper, we provide a method for aligning texts and translations based only on internal evidence. In other words, the method depends on no information about the languages involved beyond what can be derived from the texts themselves. Furthermore, the computations on which it is based are straightforward and robust. The plan rests on a relationship between word and sentence alignments arising from the observation that a pair of sentences containing an aligned pair of words must themselves be aligned. It follows that a partial alignment on the word level could induce a much more complete alignment on the sentence level.

A solution to the alignment problem consists of a subset of the Cartesian product of the sets of source and target sentences. The process starts from an initial subset excluding pairs whose relative positions in their respective texts is so different that the chance of their being aligned is extremely low. This potentially alignable set of sentences forms the basis for a relaxation process that proceeds as follows. An initial set of candidate word alignments is produced by choosing pairs of words that tend to occur in possibly aligned sentences. The idea is to propose a pair of words for alignment if they have similar distributions in their respective texts. The distributions of a pair of words are similar if most of the sentences in which the first word occurs are alignable with sentences in which the second occurs, and vice versa. The most apparently reliable of these word alignments are then used to induce a set of sentence alignments that will be a subset of the eventual result. A new estimate is now made of what sentences are alignable based on the fact that we are now committed to aligning certain pairs. Because sentence pairs are never removed from the set of alignments, the process converges to the point when no new ones can be found; then it stops.

In the next section, we describe the algorithm. In Section 3 we describe additions to the basic technique required to provide for morphology, that is, relatively superficial variations in the forms of words. In Section 4 we show the results of applying a program that embodies these techniques to an article from *Scientific American* and its German translation in *Spektrum der Wissenschaft*. In Section 5 we discuss other approaches to the alignment problem that were subsequently undertaken by other researchers (Gale and Church 1991; Brown, Lai, and Mercer 1991). Finally, in Section 6, we consider ways in which our present methods might be extended and improved.

2. The Alignment Algorithm

2.1 Data Structures

The principal data structures used in the algorithm are the following:

Word-Sentence Index (WSI). One of these is prepared for each of the texts. It is a table with an entry for each different word in the text showing the sentences in which that word occurs. For the moment, we may take a word as being simply a distinct sequence of letters. If a word occurs more than once in a sentence, that sentence occurs on the list once for each occurrence.

Alignable Sentence Table (AST). This is a table of pairs of sentences, one from each text. A pair is included in the table at the beginning of a pass if that pair is a candidate for association by the algorithm in that pass.

Word Alignment Table (WAT). This is a list of pairs of words, together with similarities and frequencies in their respective texts, that have been aligned by comparing their distributions in the texts.

Sentence Alignment Table (SAT). This is a table that records for each pair of sentences how many times the two sentences were set in correspondence by the algorithm.

Some additional data structures were used to improve performance in our implementation of the algorithm, but they are not essential to an understanding of the method as a whole.

2.2 Outline of the Algorithm

At the beginning of each cycle, an AST is produced that is expected to contain the eventual set of alignments, generally amongst others. It pairs the first and last sentences of the two texts with a small number of sentences from the beginning and end of the other text. Generally speaking, the closer a sentence is to the middle of the text, the larger the set of sentences in the other text that are possible correspondents for it.

The next step is to hypothesize a set of pairs of words that are assumed to correspond based on similarities between their distributions in the two texts. For this purpose, a word in the first text is deemed to occur at a position corresponding to a word in the second text if they occur in a pair of sentences that is a member of the AST. Similarity of distribution is a function of the number of corresponding sentences in which they occur and the total number of occurrences of each. Pairs of words are entered in the WAT if the association between them is so close that it is not likely to be the result of a random event. In our algorithm, the closeness of the association is estimated on the basis of the similarity of their distributions and the total number of occurrences.

The next step is to construct the SAT, which, in the last pass, will essentially become the output of the program as a whole. The idea here is to associate sentences that contain words paired in the WAT, giving preference to those word pairs that appear to be more reliable. Multiple associations are recorded.

If there are to be further passes of the main body of the algorithm, a new AST is then constructed in light of the associations in the SAT. Associations that are supported some minimum number of times are treated just as the first and last sentences of the texts were initially; that is, as places at which there is known to be a correspondence. Possible correspondences are provided for the intervening sentences by

the same interpolation method initially used for all sentences in the middle of the texts.

In preparation for the next pass, a new set of corresponding words is now hypothesized using distributions based on the new AST, and the cycle repeats.

2.3 The Algorithm

The main algorithm is a relaxation process that leaves at the end of each pass a new WAT and SAT, each presumably more refined than the one left at the end of the preceding pass. The input to the whole process consists only of the WSIs of the two texts. Before the first pass of the relaxation process, an initial AST is computed simply from the lengths of the two texts:

Construct Initial AST. If the texts contain m and n sentences respectively, then the table can be thought of as an $m \times n$ array of ones and zeros. The average number of sentences in the second text corresponding to a given one in the first text is n/m, and the average position of the sentence in the second text corresponding to the i-th sentence in the first text is therefore $i \cdot n/m$. In other words, the expectation is that the true correspondences will lie close to the diagonal. Empirically, sentences typically correspond one for one; correspondences of one sentence to two are much rarer, and correspondences of one to three or more, though they doubtless occur, are very rare and were unattested in our data. The maximum deviation can be stochastically modeled as $O(\sqrt{n})$, the factor by which the standard deviation of a sum of n independent and identically distributed random variables multiplies.[1]

We construct the initial AST using a function that pairs single sentences near the middle of the text with as many as $O(\sqrt{n})$ sentences in the other text; it is generously designed to admit all but the most improbable associations. Experience shows that because of this policy the results are highly insensitive to the particular function used to build this initial table.[2]

The main body of the relaxation process consists of the following steps:

Build the WAT. For all sentences s^A in the first text, each word in s^A is compared with each word in those sentences s^B of the second text that are considered as candidates for correspondence, i.e., for which $\langle s^A, s^B \rangle \in$ AST. A pair of words is entered into the WAT if the distributions of the two words in their texts are sufficiently similar and if the total number of occurrences indicates that this pair is unlikely to be the result of a spurious match. Note that the number of comparisons of the words in two sentences is quadratic only in the number of words in a sentence, which can be assumed to be not a function of the length of the text. Because of the constraint on the maximum deviation from the diagonal as outlined above, the computational complexity of the algorithm is bound by $O(n\sqrt{n})$ in each pass.

1 In such a model, each random variable would correspond to a translator's choice to move away from the diagonal in the AST by a certain distance (which is assumed to be zero mean, Gaussian distributed). However, the specific assumptions about the maximum deviation are not crucial in that the algorithm was observed to be insensitive to such modifications.

2 The final results showed that no sentence alignment is at a distance greater than ten from the diagonal in texts of 255 and 300 sentences. Clearly, any such prior knowledge could be used for a significant speed-up of the algorithm, but it was our goal to adopt as few prior assumptions as possible.

Our definition of the similarity between a pair of words is complicated by the fact that the two texts have unequal lengths and that the AST allows more than one correspondence, which means that we cannot simply take the inner product of the vector representations of the word's occurrences. Instead, we use as a measure of similarity:[3]

$$\frac{2c}{N_A(v) + N_B(w)}$$

where c is the number of corresponding positions, and $N_T(x)$ is the number of occurrences of the word x in the text T. This is essentially Dice's coefficient (Rijsbergen 1979). Technically, the value of c is the cardinality of the largest set of pairs $\langle i, j \rangle$ such that

1. $\langle s_i^A(v), s_j^B(w) \rangle \in$ AST, where $s_z^T(x)$ is the sentence in text T that contains the z-th occurrence of word x.
2. Pairs are *non-overlapping* in the sense that, if $\langle a, b \rangle$ and $\langle c, d \rangle$ are distinct members of the set then they are distinct in both components, that is, $a \neq c$ and $b \neq d$.

Suppose that the word "dog" occurs in sentences 50, 52, 75, and 200 of the English text, and "Hund" in sentences 40 and 180 of the German, and that the AST contains the pairs $\langle 50, 40 \rangle$, $\langle 52, 40 \rangle$, and $\langle 200, 180 \rangle$, among others, but not $\langle 75, 40 \rangle$. There are two sets that meet the requirements, namely $\{\langle 1,1 \rangle, \langle 4,2 \rangle\}$ and $\{\langle 2,1 \rangle, \langle 4,2 \rangle\}$. The set $\{\langle 1,1 \rangle, \langle 2,1 \rangle, \langle 4,2 \rangle\}$ is excluded on the grounds that $\langle 1,1 \rangle$ and $\langle 2,1 \rangle$ overlap in the above sense—the first occurrence of "Hund" is represented twice. In the example, the similarity would be computed as $\frac{2}{4+2-2} = \frac{1}{2}$, regardless of the ambiguity between $\langle 1,1 \rangle$ and $\langle 2,1 \rangle$.

The result of the comparisons of the words in all of the sentences of one text with those in the other text is that the word pairs with the highest similarity are located. Comparing the words in a sentence of one text with those in a sentence of the other text carries with it an amortized cost of constant computational complexity,[4] if the usual memory-processing tradeoff on serial machines is exploited by maintaining redundant data structures such as multiple hash tables and ordered indexed trees.[5]

The next task is to determine for each word pair, whether it will actually be entered into the WAT: the WAT is a sorted table where the more reliable pairs are put before less reliable ones. For this purpose, each entry contains, as well as the pair of words themselves, the frequencies of those words in their respective texts and the similarity between them. The closeness of the association between two words, and thus their rank in the WAT, is evaluated with respect to their similarity and the total number of their occurrences. To understand why similarity cannot be used

3 Throughout this paper, we use the word *similarity* to denote this similarity measure, which does not necessarily have to be an indicator of what one would intuitively describe as "similar" words. In particular, we will later see that similarity alone, without consideration of the total frequency, is not a good indicator for "similarity."

4 The basic idea is this: more processing has to be done to compute the similarity of a high-frequency word to another frequent word, but there are also more places at which this comparison can later be saved. Recall also that we assume sentence length to be independent of text length.

5 For very large corpora, this might not be feasible. However, large texts can almost invariably be broken into smaller pieces at natural and reliable places, such as chapter and section headings.

alone, note that there are far more one-frequency words than words of higher frequency. Thus, a pair of words with a similarity of 1, each of them occurring only once, may well be the result of a random event. If such a pair was proposed for entry into the WAT, it should only be added with a low priority.

The exact stochastic relation is depicted in Figure 1, where the probability is shown that a word of a frequency k that was aligned with a word in the other text with a certain similarity s is just the result of a random process.[6] Note that, for a high-frequency word that has a high similarity with some other word (right front corner), it is very unlikely (negligible plateau height) that this association has to be attributed to chance. On the other hand, low similarities (back) can easily be attained by just associating arbitrary words. Low-frequency words—because there are so many of them in a text—can also achieve a high similarity with some other words without having to be related in an interesting way. This can be intuitively explained by the fact that the similarity of a high-frequency word is based on a pattern made up of a large number of instances. It is therefore a pattern that is unlikely to be replicated by chance. Furthermore, since there are relatively few high-frequency words, and they can only contract high similarities with other high-frequency words, the number of possible correspondents for them is lower, and the chance of spurious associations is therefore less on these grounds also. Note that low-frequency words with low similarity (back left corner) have also a low probability of being spuriously associated to some other word. This is because low-frequency words can achieve a low similarity only with words of a high frequency, which in turn are rare in a text, and are therefore unlikely to be associated spuriously.[7]

Our algorithm does not use all the detail in Figure 1, but only a simple discrete heuristic: a word pair whose similarity exceeds some threshold is assigned to one of two or three segments of the WAT, depending on the word frequency. A segment with words of higher frequency is preferred to lower-frequency segments. Within each segment, the entries are sorted in order of decreasing similarity and, in case of equal similarities, in order of decreasing frequency. In terms of Figure 1, we take a rectangle from the right front. We place the left boundary as far to the left as possible, because this is where most of the words are.

Build the SAT. In this step, the correspondences in the WAT are used to establish a mapping between sentences of the two texts. In general, these new

6 The basis for this graph is an analytic derivation of the probability that a word with a certain frequency in a 300-sentence text matches some random pattern with a particular similarity. The analytic formula relies on word-frequency data derived from a large corpus instead of on a stochastic model for word frequency distribution (such as Zipf's law, which states that the frequency with which words occur in a text is indirectly proportional to the number of words with this frequency; for a recent discussion of more accurate models, see also Baayen [1991]). Clearly, the figure is dependent on the state of the AST (e.g. lower similarities become more acceptable as the AST becomes more and more narrow), but the thresholds relevant to our algorithm can be precomputed at compile-time. The figure shown would be appropriate to pass 3 in our experiment. In the formula used, there are a few reasonable simplifications concerning the nature of the AST; however, a Monte-Carlo simulation that is exactly in accordance with our algorithm confirmed the depicted figure in every essential detail.

7 This discussion could also be cast in an information theoretic framework using the notion of "mutual information" (Fano 1961), estimating the variance of the degree of match in order to find a frequency-threshold (see Church and Hanks 1990).

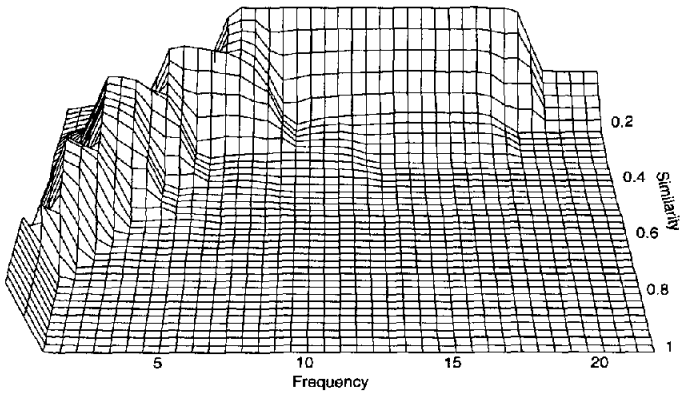

Figure 1
Likelihood that a word pair is a spurious match as a function of a word's frequency and its similarity with a word in the other text (maximum 0.94).

associations are added to the ones inherited from the preceding pass. It is an obvious requirement of the mapping that lines of association should not cross. At the beginning of the relaxation process, the SAT is initialized such that the first sentences of the two texts, and the last sentences, are set in correspondence with one another, regardless of any words they may contain. The process that adds the remaining associations scans the WAT in order and applies a three-part process to each pair $\langle v, w \rangle$.

1. Construct the *correspondence set* for $\langle v, w \rangle$ using essentially the same procedure as in the calculation of the denominator, c, of word similarities above. Now, however, we are concerned to avoid ambiguous pairs as characterized above. The set contains a sentence pair $\langle s_i^A(v), s_j^B(w) \rangle$ if (1) $\langle s_i^A(v), s_j^B(w) \rangle \in$ AST, and (2) v occurs in no other sentence h (resp. w in no g) such that $\langle s_i^A(v), h \rangle$ (resp. $\langle g, s_j^B(w) \rangle$) is also in the AST.
2. If any sentence pair in the correspondence set crosses any of the associations that have already been added to the SAT, the word pair is rejected as a whole. In other words, if a given pair of sentences correspond, then sentences preceding the first of them can be associated only with sentences preceding the second.
3. Add each sentence pair in the correspondence set of the word pair $\langle v, w \rangle$ to the SAT. A count is recorded of the number of times a particular association is supported. These counts are later thresholded when a new AST is computed or when the process terminates.

Build a New AST. If there is to be another pass of the relaxation algorithm, a new AST must be constructed as input to it. This is based on the current SAT and is derived from it by supplying associations for sentences for which it

provides none. The idea is to fill gaps between associated pairs of sentences in the same manner that the gap between the first and the last sentence was filled before the first pass. However, only sentence associations that are represented more than some minimum number of times in the SAT are transferred to the AST. In what follows, we will refer to these sentence pairs as *anchors*.

As before, it is convenient to think of the AST as a rectangular array, even though it is represented more economically in the program. Consider a maximal sequence of empty AST entries, that is, a sequence of sentences in one text for which there are no associated sentences in the other, but which is bounded above and below by an anchor. The new associations that are added lie on and adjacent to the diagonal joining the two anchors. The distance from the diagonal is a function of the distance of the current candidate sentence pair and the nearest anchor. The function is the same one used in the construction of the initial AST.

Repeat. Build a new WAT and continue.

3. Morphology

As we said earlier, the basic alignment algorithm treats words as atoms; that is, it treats strings as instances of the same word if they consist of identical sequences of letters, and otherwise as totally different. The effect of this is that morphological variants of a word are not seen as related to one another. This might not be seen as a disadvantage in all circumstances. For example, nouns and verbs in one text might be expected to map onto nouns with the same number and verbs with the same tense much of the time. But this is not always the case and, more importantly, some languages make morphological distinctions that are absent in the other. German, for example, makes a number of case distinctions, especially in adjectives, that are not reflected in the morphology of English. For these reasons, it seems desirable to allow words to contract associations with other words both in the form in which they actually occur, and in a more normalized form that will throw them together with morphologically related other words in the text.

3.1 The Basic Idea

The strategy we adopted was to make entries in the WSI, not only for maximal strings of alphabetic characters occurring in the texts, but also for other strings that could usefully be regarded as normalized forms of these.

Clearly, one way to obtain normalized forms of words is to employ a fully fledged morphological analyzer for each of the languages. However, we were concerned that our methods should be as independent as possible of any specific facts about the languages being treated, since this would make them more readily usable. Furthermore, since our methods attend only to very gross features of the texts, it seemed unreasonable that their success should turn on a very fine analysis at any level. We argue that, by adding a guess as to how a word should be normalized to the WSI, we remove no associations that could have been formed on the basis of the original word, but only introduce the possibility of some additional associations. Also, it is unlikely that an incorrect normalization will contract any associations at all, especially in view of the fact that these forms, because they normalize several original forms, tend to occur more often. They will therefore rarely be misleading.

For us, a normalized form of a word is always an initial or a final substring of that word—no attention is paid to morphographemic or word-internal changes. A word is broken into two parts, one of which becomes the normalized form, if there is evidence that the resulting prefix and suffix belong to a paradigm. In particular, both must occur as prefixes and suffixes of other forms.

3.2 The Algorithm

The algorithm proceeds in two stages. First a data structure, called the *trie*, is constructed in which information about the occurrences of potential prefixes and suffixes in the text is stored. Second, words are split, where the trie provides evidence for doing so, and one of the resulting parts is chosen as the normalization.

1. A trie (Knuth 1973; pp. 481–490) is a data structure for associating information with strings of characters. It is particularly economical in situations where many of the strings of interest are substrings of others in the set. A trie is in fact a tree, with a branch at the root node for every character that begins a string in the set. To look up a string, one starts at the root, and follows the branch corresponding to its first character to another node. From there, the branch for the second character is followed to a third node, and so on, until either the whole string has been matched, or it has been discovered not to be in the set. If it is in the set, then the node reached after matching its last character contains whatever information the structure contains for it. The economy of the scheme lies in the fact that a node containing information about a string also serves as a point on the way to longer strings of which the given one is a prefix. In this application, two items of information are stored with a string, namely the number of textual words in which it occurs as a prefix and as a suffix.

2. Consider the possibility of breaking an n-letter word before the i-th character of the word ($1 < i \leq n$). The conditions for a break are: The number of other words starting with characters $1 \cdots i-1$ of the current word must be greater than the number of words starting with characters $1 \cdots i$ because, if the characters $1 \cdots i-1$ constitute a useful prefix, then this prefix must be followed, in different words, by other suffixes than characters $i \cdots n$. So, consider the word "wanting," and suppose that we are considering the possibility of breaking it before the 5th character, "i." For this to be desirable, there must be other words in the text, such as "wants," and "wanted," that share the first $i-1=4$ characters. Conversely, there must be more words ending with characters $i \cdots n$ of the word than with $i-1 \cdots n$. So, there must be more words with the suffix "ing" than with the suffix "ting"; for example "seeing" and "believing."

   There is a function from potential break points in words to numbers whose value is maximized to choose the best point at which to break. If p and s are the potential prefix and suffix, respectively, and $P(p)$ and $S(s)$ are the number of words in the text in which they occur as such, the value of the function is $kP(p)S(s)$. The quantity k is introduced to enable us to prefer certain kinds of breaks over others. For the English and German texts used in our experiments, $k = length(p)$ so as to favor long prefixes on the grounds that both languages are primarily suffixing. If

the function has the same value for more than one potential break point, the one farthest to the right is preferred, also for the reason that we prefer to maximize the lengths of prefixes.

Once it has been decided to divide a word, and at what place, one of the two parts is selected as the putative canonical form of the word, namely, whichever is longer, and the prefix if both are of equal length. Finally, any other words in the same text that share the chosen prefix (suffix) are split at the corresponding place, and so assigned to the same canonical form.

The morphological algorithm treats words that appear hyphenated in the text specially. The hyphenated word is treated as a unit, just as it appears, and so are the strings that result from breaking the word at the hyphens. In addition, the analysis procedure described above is applied to these components, and any putative normal forms found are also used. It is worth pointing out that we received more help from hyphens than one might normally expect in our analysis of the German texts because of a tendency on the part of the *Spektrum der Wissenschaft* translators, following standard practice for technical writing, of hyphenating compounds.

4. Experimental Results

In this section, we show some of the results of our experiments with these algorithms, and also data produced at some of the intermediate stages. We applied the methods described here to two pairs of articles from *Scientific American* and their German translations in *Spektrum der Wissenschaft* (see references). The English and German articles about human-powered flight had 214 and 162 sentences, respectively; the ones about cosmic rays contained 255 and 300 sentences, respectively. The first pair was primarily used to develop the algorithm and to determine the various parameters of the program. The performance of the algorithm was finally tested on the latter pair of articles. We chose these journals because of a general impression that the translations were of very high quality and were sufficiently "free" to be a substantial challenge for the algorithm. Furthermore, we expected technical translators to adhere to a narrow view of semantic accuracy in their work, and to rate the importance of this above stylistic considerations. Later we also give results for another application of our algorithm to a larger text of 1257 sentences that was put together from two days from the French–English Hansard corpus.

Table 1 shows the first 50 entries of the WAT after pass 1 of the algorithm. It shows part of the first section of the WAT (lines 1–23) and the beginning of the second (lines 24–50). The first segment contains words or normalized forms with more than 7 occurrences and a similarity not less than 0.8. Strings shown with a following hyphen are prefixes arising from the morphological procedure; strings with an initial hyphen are suffixes. Naturally, some of the word divisions are made in places that do not accurately reflect linguistic facts. For example, English "proto-" (1) comes from "proton" and "protons"; German "-eilchen" (17) is the normalization for words ending in "-teilchen" and, in the same way, "-eistung" (47) comes from "-leistung."

Of these 50 word pairs, 42 have essentially the same meanings. We take it that "erg" and "Joule," in line 4, mean the same, *modulo* a change in units. Also, it is not unreasonable to associate pairs like "primary"/"sekundären" (26) and "electric"/"Feld" (43), on the grounds that they tend to be used together. The pair "rapid-"/"Pulsare-" (49) is made because a pulsar is a rapidly spinning neutron star and some such phrase

Table 1
The WAT after pass 1.

	English	German	Eng. Freq.	Similarity
1	proto-	Proto-	14	1
2	proton-	Proton-	13	1
3	interstellar	interstellare-	12	1
4	ergs	Joule	10	1
5	electric-	elektrisch-	9	1
6	pulsar-	Pulsar-	17	16/17
7	photo-	Photo-	14	14/15
8	and	und	69	11/12
9	per	pro	12	11/12
10	relativ-	relativ-	11	10/11
11	atmospher-	Atmosphäre-	10	10/11
12	Cygnus	Cygnus	63	59/65
13	cosmic-	kosmische-	81	39/43
14	volts	Elektronenvolt	19	19/21
15	telescope-	Teleskop-	9	8/9
16	univers-	Univers-	8	7/8
17	particle-	-eilchen	53	51/59
18	shower-	Luftschauer-	20	19/22
19	X-ray-	Röntgen-	19	19/22
20	electrons	Elektronen	12	11/13
21	source-	Quelle-	40	37/45
22	magnetic	Magnetfeld	11	9/11
23	ray-	Strahlung-	141	135/167
24	Observatory	diesem	6	1
25	shower	Gammaquant	6	1
26	primary	sekundären	6	1
27	percent	Prozent	6	1
28	galaxies	Galaxien	5	1
29	Crimean	Krim	5	1
30	ultrahigh-	ultraho-	5	1
31	density	Dichte	5	1
32	synchrotron	Synchrotronstrahlung	5	1
33	activ-	aktiv-	5	1
34	supernova	Supernova-Explosion-	5	1
35	composition	Zusammensetzung	5	1
36	detectors	primäre-	5	1
37	data	Daten-	7	7/8
38	University	Universit-	7	6/7
39	element-	-usammensetzung	7	6/7
40	neutron	Neutronenstern	7	6/7
41	Cerenkov	Cerenkov-Licht-	7	6/7
42	spinning	rotier-	6	6/7
43	electric	Feld	6	5/6
44	lines	-inien	6	5/6
45	medium	Medium	6	5/6
46	estimate-	abschätz-	6	5/6
47	output	-eistung	6	5/6
48	bright-	Astronom-	5	5/6
49	rapid-	Pulsare-	5	5/6
50	proposed	vorgeschlagen	6	5/6

occurs with it five out of six times. Notice, however, that the association "pulsar-" "Pulsar-" is also in table (6). Furthermore, the German strings "Pulsar" and "Pulsar-" are both given correct associations in the next pass (lines 17 and 20 of Table 2).

The table shows two interesting effects of the morphological analysis procedure. The word "shower" is wrongly associated with the word "Gammaquant" (25) with a frequency of 6, but the prefix "shower-" is correctly associated with "Luftschauer-"

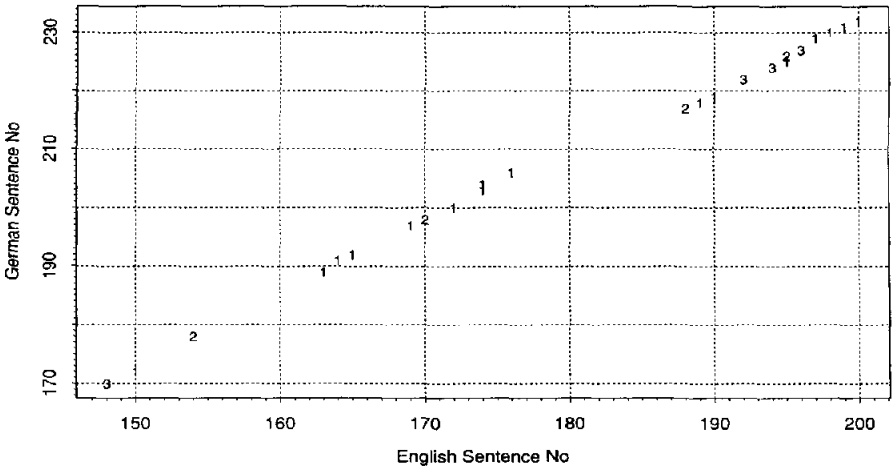

Figure 2
The SAT after pass 1.

(18) with a frequency of 20. On the other hand, the incorrect association of "element" with "-usammensetzung" (39) is on the basis of a normalized form (for words ending in "Zusammensetzung"), whereas "Zusammensetzung," unnormalized, is correctly associated with "composition" (35). Totally unrelated words are associated in a few instances, as in "Observatory"/"diesem" (24), "detectors"/"primäre-" (36), and "bright-"/"Astronom-" (48). Of these only the second remains at the end of the third pass. The English "Observatory" is then properly associated with the German word "Observatorium-." At that stage, "bright-" has no association.

Figure 2 shows part of the SAT at the end of pass 1 of the relaxation cycle. Sentences in the English text and in the German text are identified by numbers on the abscissa and the ordinate respectively. Entries in the array indicate that the sentences are considered to correspond. The numbers show how often a particular association is supported, which is essentially equivalent to how many word pairs in the WAT support such an association. If there are no such numbers, then no associations have been found for it at this stage. For example, the association of English sentence 148 with German sentence 170 is supported by three different word pairs. It is already very striking how strongly occupied entries in this table constrain the possible entries in the unoccupied slots.

Figure 3 shows part of the AST before pass 2. This is derived directly from the material illustrated in Figure 2. The abscissa gives the English sentence number and in direction of the ordinate the associated German sentences are shown (bullet). Those sentence pairs in Figure 2 supported by at least three word pairs, namely those shown on lines 148, 192, 194, and 196, are assumed to be reliable, and they are the only associations shown for these sentences in Figure 3. Candidate associations have been provided for the intervening sentences by the interpolation method described above. Notice that the greatest number of candidates are shown against sentences occurring midway between a pair assumed to have been reliably connected (English sentence numbers 169 to 171).

Table 2 shows the first 100 entries of the WAT after pass 3, where the threshold

Table 2
The WAT after pass 3.

	English	German	Eng. Freq.	Similarity
1	interstellar	interstellare-	12	1
2	ergs	Joule	10	1
3	per	pro	12	11/12
4	univers-	Univers-	8	7/8
5	proto-	Proto-	14	13/15
6	X-ray-	Röntgen-	19	19/22
7	proton-	Proton-	13	6/7
8	volts	Elektronenvolt	19	9/11
9	photo-	Photo-	14	13/16
10	light-	Licht-	23	21/26
11	earth	Erde	9	4/5
12	accelerate-	beschleunigt	9	7/9
13	object	Objekt	9	7/9
14	Cygnus	Cygnus	63	27/35
15	accelerat-	beschleunig-	18	16/21
16	model-	Modell-	17	16/21
17	pulsars	Pulsare-	8	3/4
18	cosmic-	kosmische-	81	35/47
19	galaxy	Milchstraße	19	17/23
20	pulsar-	Pulsar-	17	14/19
21	electrons	Elektronen	12	5/7
22	magnetic	Magnetfeld-	11	5/7
23	shower-	Luftschauer-	20	17/24
24	telescope-	Teleskop-	9	7/10
25	source-	Quelle-	40	33/49
26	Second-	Sekund-	20	2/3
27	low-	nied-	9	2/3
28	part-	Teil-	59	49/76
29	and	und	69	9/14
30	electric-	elektrisch-	9	7/11
31	gamma-	Gammastrahl-	61	27/43
32	gas-	Gas-	16	5/8
33	relativ-	relativ-	11	8/13
34	atmospher-	Atmosphäre-	10	8/13
35	direction	-ichtung	10	3/5
36	years	Jahre-	11	10/17
37	object-	Objekt-	14	10/17
38	period-	Stunden-	11	7/12
39	electro-	elektr-	83	63/109
40	only	Nur	18	15/26
41	source	-uelle	26	4/7
42	photon-	Photon-	10	4/7
43	high-energy	hochenerg-	13	9/16
44	directions	-ichtungen	8	5/9
45	thousand-	Tausend-	8	5/9
46	stars	Sterne-	11	6/11
47	number	Anzahl	8	6/11
48	interact-	wechselwirk-	9	6/11
49	signal	Signal-	12	7/13
50	the	die-	496	313/582
51	energy	Energie	28	22/41
52	wave-	Wellen-	13	8/15
53	star-	Stern-	29	9/17
54	sources	Quellen	14	11/21
55	nucle-	Atom-	19	12/23
56	of	ein-	304	1/2
57	not	nicht	30	1/2
58	ray	Gammaquant-	14	1/2
59	arrival	Ankunfts-	9	1/2

Table 2
(continued) The WAT after pass 3.

60	percent	Prozent	6	1
61	ultrahigh-	ultraho-	5	1
62	galaxies	Galaxien	5	1
63	composition	Zusammensetzung	5	1
64	Crimean	Krim	5	1
65	supernova	Supernova-Explosion-	5	1
66	activ-	aktiv-	5	1
67	synchrotron	Synchrotronstrahlung	5	1
68	detectors	primäre-	5	1
69	muons	Myonen	4	1
70	massive	Masse-	4	1
71	meteorite-	Meteorit-	4	1
72	Low-energy	niederenergetische-	4	1
73	Fermi	Fermi-	4	1
74	decay-	Zerfall-	4	1
75	discovery	Entdeckung	4	1
76	limit	Grenze	4	1
77	ground	Erdboden	4	1
78	day-	Tag-	3	1
79	Robert	Robert	3	1
80	mirrors	Spiegel-	3	1
81	absorption	Absorptionslinie-	3	1
82	David	David	3	1
83	average	Mittel-	3	1
84	light-years	Lichtjahre	3	1
85	Neutrons	Neutronen	3	1
86	Gregory-	Gregory-	3	1
87	explosions	Supernova-Explosionen	3	1
88	electrically	elektrisch	3	1
89	electromagnetic	elektromagnetische-	3	1
90	candidates	Kandidaten	3	1
91	data	Daten-	7	7/8
92	University	Universit-	7	6/7
93	spinning	rotier-	6	6/7
94	neutron	Neutronenstern	7	6/7
95	proposed	vorgeschlagen	6	5/6
96	lines	-inien	6	5/6
97	colleague-	Kollegen	4	4/5
98	interactions	Wechselwirkungen	5	4/5
99	Physic-	Physik-	5	4/5
100	models	Modelle-	4	4/5

for the similarity was lowered to 0.5. As we pointed out earlier, most of the incorrect associations in Table 1 have been eliminated. German "Milchstraße" (19) is not a translation of the English "galaxy," but the Milky Way is indeed a galaxy and "the galaxy" is sometimes used in place of "Milky Way" where the reference is clear. The association between "period-" and "Stunden-" (38) is of a similar kind. The words are strongly associated because of recurring phrases of the form "in a 4.8-hour period."

Figure 4 gives the SAT after pass 3. It is immediately apparent, first, that the majority of the sentences have been associated with probable translations and, second, that many of these associations are very strongly supported. For example, note that the correspondence between English sentence 190 and German sentence 219 is supported 21 times. Using this table, it is in fact possible to locate the translation of a given English sentence to within two or three sentences in the German text, and usually more closely than that. However, some ambiguities remain. Some of the apparent anomalies come

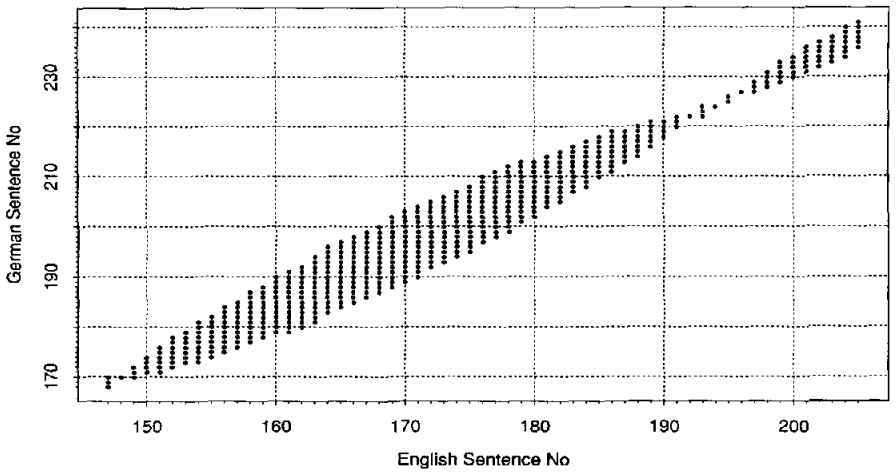

Figure 3
The AST before pass 2.

from stylistic differences in the way the texts were presented in the two journals. The practice of *Scientific American* is to collect sequences of paragraphs into a logical unit by beginning the first of them with an oversized letter. This is not done in *Spektrum der Wissenschaft*, which instead provides a subheading at these points. This therefore appears as an insertion in the translation. Two such are sentences number 179 and 233, but our procedure has not created incorrect associations for them.

Recall that the alignment problem derives its interest from the fact that single sentences are sometimes translated as sequences of sentences and conversely. These cases generally stand out strongly in the output that our method delivers. For example, the English sentence pair (5, 6):

> Yet whereas many of the most exciting advances in astronomy have come from the detailed analysis of X-ray and radio sources, until recently the source of cosmic rays was largely a matter of speculation. They seem to come from everywhere, raining down on the earth from all directions at a uniform rate.

is rendered in German by the single sentence (5):

> Dennoch blieben die Quellen der kosmischen Strahlung, die aus allen Richtungen gleichmäßig auf die Erde zu treffen scheint, bis vor kurzem reine Spekulation, während einige der aufregendsten Fortschritte in der Astronomie aus dem detaillierten Studium von Röntgen- und Radiowellen herrührten.

The second English sentence becomes a relative clause in the German.

More complex associations also show up clearly in the results. For example, English sentences 218 and 219 are translated by German sentences 253, 254, and 255,

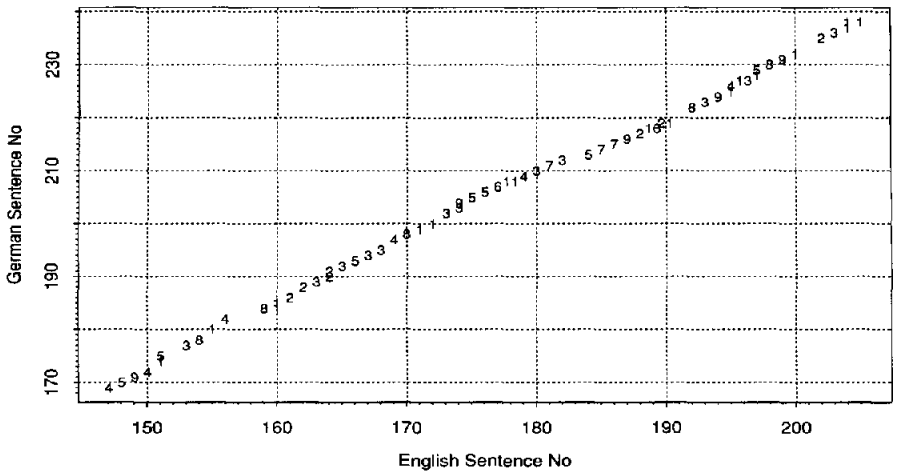

Figure 4
The SAT after pass 3.

where 254 is a translation of the latter part of 218 and the early part of 219:

> When a proton strikes a gas nucleus, it produces three kinds of pion, of which one kind decays into two gamma rays. The gamma rays travel close to the original trajectory of the proton, and the model predicts they will be beamed toward the earth at just two points on the pulsars orbit around the companion star.

> Trifft ein Proton auf einen Atomkern in dieser Gashülle, werden drei Arten von Pionen erzeugt. Die neutralen Pionen zerfallen in jeweils zwei Gammaquanten, die sich beinahe in dieselbe Richtung wie das ursprüngliche Proton bewegen. Nach der Modellvorstellung gibt es gerade zwei Positionen im Umlauf des Pulsars um seinen Begleitstern, bei denen die Strahlung in Richtung zum Beobachter auf der Erde ausgesandt wird.

Another example is provided by English sentences 19 and 20, which appear in German as sentences 21 and 22. However the latter part of English sentence 19 is in fact transferred to sentence 22 in the German. This is also unmistakable in the final results. Notice also, in this example, that the definition of "photon" has become a parenthetical expression at the beginning of the second German sentence, a fact which is not reflected.

> The other end of the cosmic-ray energy spectrum is defined somewhat arbitrarily: any quantum greater than 10^8 electron volts arriving from space is considered a cosmic ray. The definition encompasses not only particles but also gamma-ray photons, which are quanta of electromagnetic radiation.

Table 3
Correctness of sentence alignment in the various passes of the algorithm.

Pass	Correctness in SAT	Coverage of SAT	Constraint by AST
1	100 %	12 %	4 %
2	100 %	47 %	17 %
3	100 %	89 %	38 %
4	99.7 %	96 %	41 %

> Das untere Ende des Spektrums der kosmischen Strahlen ist verhältnismäßig unscharf definiert. Jedes Photon (Quant der elektromagnetischen Strahlung) oder Teilchen mit einer Energie von mehr als 10^8 Elektronenvolt, das aus dem Weltraum eintrifft, bezeichnet man als kosmischen Strahl.

It frequently occurred in our data that sentences that were separated by colons or semicolons in the original appeared as completely distinct sentences in the German translation. Indeed, the common usage in the two languages would probably have been better represented if we had treated colons and semicolons as sentence separators, along with periods, question marks, and the like. There are, of course, situations in English in which these punctuation marks are used in other ways, but they are considerably less frequent and, in any case, it seems that our program would almost always make the right associations. An example involving the colon is to be found in sentence 142 of the original, translated as sentences 163 and 164:

> The absorption lines established a lower limit on the distance of Cygnus X-3: it must be more distant than the farthest hydrogen cloud, which is believed to lie about 37,000 light-years away, near the edge of the galaxy.

> Aus dieser Absorptionslinie kann man eine untere Grenze der Entfernung von Cygnus X bestimmen. Die Quelle muß jenseits der am weitesten entfernten Wasserstoff-Wolke sein, also weiter als ungefähr 37000 Lichtjahre entfernt, am Rande der Milchstraße.

English sentence 197, containing a semicolon, is translated by German sentences 228 and 229:

> The estimate is conservative; because it is based on the gamma rays observed arriving at the earth, it does not take into account the likelihood that Cygnus X emits cosmic rays in all directions.

> Dies ist eine vorsichtige Abschätzung. Sie ist nur aus den Gammastrahlen-Daten abgeleitet, die auf der Erde gemessen werden; daß Cygnus X-3 wahrscheinlich kosmische Strahlung in alle Richtungen aussendet, ist dabei noch nicht berücksichtigt.

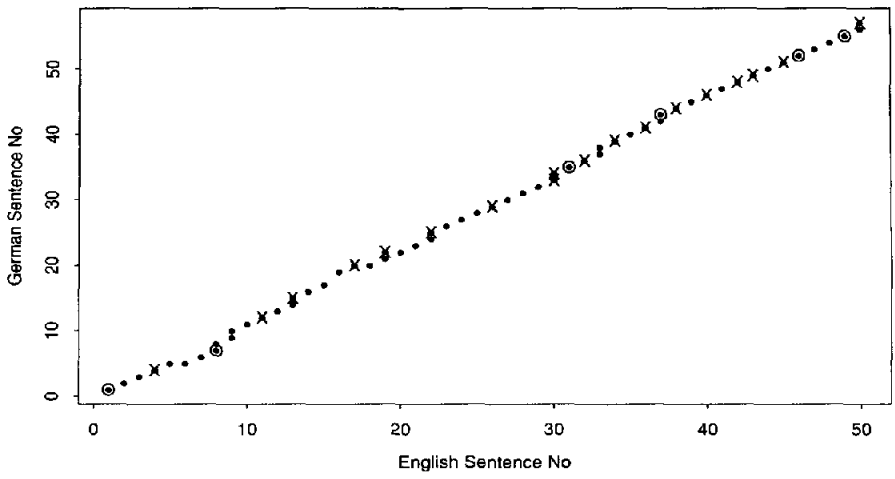

Figure 5
Sentence alignment of the first 50 sentences of the test texts: true alignment (dots) and hypothesis of the SAT after the first pass (circles) and after the second pass (crosses).

Table 3 summarizes the accuracy of the algorithm as a function of the number of passes. The (thresholded) SAT is evaluated by two criteria: the number of correct alignments divided by the total number of alignments, and—since the SAT does not necessarily give an alignment for every sentence—the coverage, i.e., the number of sentences with at least one entry relative to the total number of sentences. An alignment is said to be correct if the SAT contains exactly the numbers of the sentences that are complete or partial translations of the original sentence. The coverage of 96% of the SAT in pass 4 is as much as one would expect, since the remaining nonaligned sentences are one-zero alignments, most of them due to the German subheadings that are not part of the English version. The table also shows that the AST always provides a significant number of candidates for alignment with each sentence before a pass: the fourth column gives the number of true sentence alignments relative to the total number of candidates in the AST. Recall that the final alignment is always a subset of the hypotheses in the AST in every preceding pass.

Figure 5 shows the true sentence alignment for the first 50 sentences (dots), and how the algorithm discovered them: in the first pass, only a few sentences are set into correspondence (circles); after the second pass (crosses) already almost half of the correspondences are found. Note that there are no wrong alignments in the first two passes. In the third pass, almost all of the remaining alignments are found (for the first 50 sentences in the figure: all), and a final pass usually completes the alignment.

Our algorithm produces very favorable results when allowed to converge gradually. Processing time in the original LISP implementation was high, typically several hours for each pass. By trading CPU time for memory massively, the time needed by a C++ implementation on a Sun 4/75 was reduced to 1.7 min for the first pass, 0.8 min for the second, and 0.5 min for the third pass in an application to this pair of articles. (Initialization, i.e., reading the files and building up the data structures, takes another 0.6 min in the beginning.) It should be noted that a naive implementation of

the algorithm without using the appropriate data structures can easily lead to times that are a factor of 30 higher and do not scale up to larger texts.

The application of our method to a text that we put together from the Hansard corpus had essentially no problem in identifying the correct sentence alignment in a process of five passes. The alignments for the first 1000 sentences of the English text were checked by hand, and seven errors were found; five of them occurred in sentences where sentence boundaries were not correctly identified by the program because of periods that did not mark a sentence boundary and were not identified as such by a very simple preprocessing program. The other two errors involved two short sentences for which the SAT did not give an alignment. Processing time increased essentially linearly (per pass): the first pass took 8.3 min, the second 3.2 min, and it further decreased until the last pass, which took 2.1 min. (Initialization took 4.2 min.) Note that the error rate depends crucially on the kind of "annealing schedule" used: if the thresholds that allow a word pair in the WAT to influence the SAT are lowered too fast, only a few passes are needed, but accuracy deteriorates. For example, in an application where the process terminated after only three passes, the accuracy was only in the eighties (estimated on the basis of the first 120 sentences of the English Hansard text checked by hand). Since processing time after the first pass is usually already considerably lower, we have found that a high accuracy can safely be attained when more passes are allowed than are actually necessary.

In order to evaluate the sensitivity of the algorithm to the lengths of the texts that are to be aligned, we applied it to text samples that ranged in length from 10 to 1000 sentences, and examined the accuracy of the WAT after the first pass; that is, more precisely, the number of word pairs in the WAT that are valid translations relative to the total number of word pairs with a similarity of not less than 0.7 (the measurements are cross-validated over different texts). The result is that this accuracy increases asymptotically to 1 with the text length, and is already higher than 80% for a text length of 100 sentences (which is sufficient to reach an almost perfect alignment in the end). Roughly speaking, the accuracy is almost 1 for texts longer than 150 sentences, and around 0.5 for text length in the lower range from 20 to 60. In other words, texts of a length of more than 150 sentences are suitable to be processed in this way; text fragments shorter than 80 sentences do not have a high proportion of correct word pairs in the first WAT, but further experiments showed that the final alignment for texts of this length is, on average, again almost perfect: the drawback of a less accurate initial WAT is apparently largely compensated for by the fact that the AST is also narrower for these texts; however, the variance in the alignment accuracies is significantly higher.

5. Related Work

Since we addressed the text translation alignment problem in 1988, a number of researchers, among them Gale and Church (1991) and Brown, Lai, and Mercer (1991), have worked on the problem. Both methods are based on the observation that the length of text unit is highly correlated to the length of the translation of this unit, no matter whether length is measured in number of words or in number of characters (see Figure 6). Consequently, they are both easier to implement than ours, though not necessarily more efficient. The method of Brown, Lai, and Mercer (1991) is based on a hidden Markov model for the generation of aligned pairs of corpora, whose parameters are estimated from a large text. For an application of this method to the Canadian Hansard, good results are reported. However, the problem was also considerably facilitated by the way the implementation made use of Hansard-specific comments

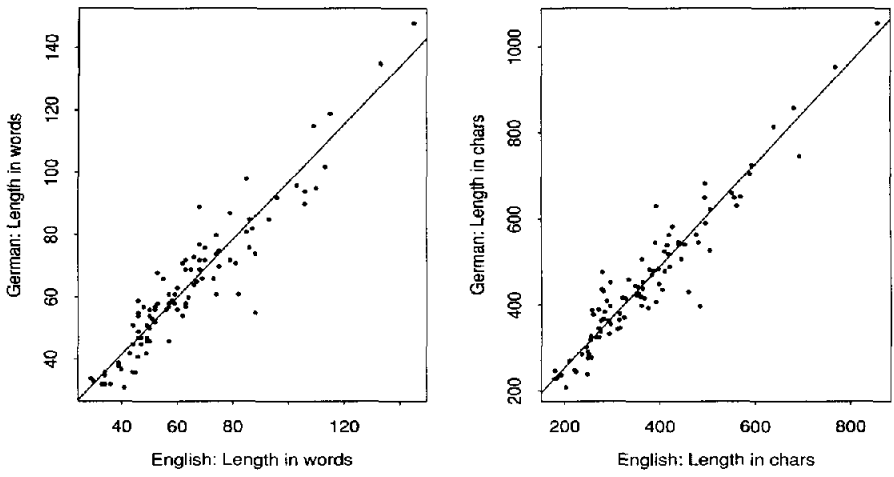

Figure 6
Lengths of Aligned Paragraphs are Correlated: Robust regression between lengths of aligned paragraphs. Left: length measured in words. Right: length measured in characters.

and annotations: these are used in a preprocessing step to find anchors for sentence alignment such that, on average, there are only ten sentences in between. Moreover, this particular corpus is well known for the near literalness of its translations, and it is therefore unclear to what extent the good results are due to the relative ease of the problem. This would be an important consideration when comparing various algorithms; when the algorithms are actually applied, it is clearly very desirable to incorporate as much prior knowledge (say, on potential anchors) as possible. Moreover, long texts can almost always be expected to contain natural anchors, such as chapter and section headings, at which to make an *a priori* segmentation.

Gale and Church (1991) note that their method performed considerably better when lengths of sentences were measured in number of characters instead of in number of words. Their method is based on a probabilistic model of the distance between two sentences, and a dynamic programming algorithm is used to minimize the total distance between aligned units. Their implementation assumes that each character in one language gives rise to, on average, one character in the other language.[8] In our texts, one character in English on average gives rise to somewhat more than 1.2 characters in German, and the correlation between the lengths (in characters) of aligned paragraphs in the two languages was with 0.952 lower than the 0.991 that are mentioned in Gale and Church (1991), which supports our impression that the *Scientific American* texts we used are hard texts to align, but it is not clear to what extent this would deteriorate the results. In applications to economic reports from the Union Bank of Switzerland, the method performs very well on simple alignments (one-to-one, one-to-two), but has at the moment problems with complex matches. The method has the

8 Recall that, in a similar way, we assumed in our implementation that one sentence in one language gives rise to, on average, n/m sentences in the other language (see first footnote in Section 2.3).

advantage of associating a score with pairs of sentences so that it is easy to extract a subset for which there is a high likelihood that the alignments are correct.

Given the simplicity of the methods proposed by Brown, Lai, and Mercer and Gale and Church, either of them could be used as a heuristic in the construction of the initial AST in our algorithm. In the current version, the number of candidate sentence pairs that are considered in the first pass near the middle of a text contributes disproportionally to the cost of the computation. In fact, as we remarked earlier, the complexity of this step is $O(n\sqrt{n})$. The proposed modification would effectively make it linear.

6. Future Work

For most practical purposes, the alignment algorithm we have described produces very satisfactory results, even when applied to relatively free translations. There are doubtless many places in which the algorithm itself could be improved. For example, it is clear that the present method of building the SAT favors associations between long sentences, and this is not surprising, because there is more information in long sentences. But we have not investigated the extent of this bias and we do not therefore know it as appropriate.

The present algorithm rests on being able to identify one-to-one associations between certain words, notably technical terms and proper names. It is clear from a brief inspection of Table 2 that very few correspondences are noticed among everyday words and, when they are, it is usually because those words also have precise technical uses. The very few exceptions include "only"/"nur" and "the"/"die-." The pair "per"/"pro" might also qualify, but if the languages afford any example of a scientific preposition, this is surely it. The most interesting further developments would be in the direction of loosening up this dependence on one-to-one associations both because this would present a very significant challenge and also because we are convinced that our present method identifies essentially all the significant one-to-one associations.

There are two obvious kinds of looser associations that could be investigated. One would consist of connections between a single vocabulary item in one language and two or more in the other, or even between several items in one language and several in the other. The other would involve connections—one–one, one–many, or many–many—between phrases or recurring sequences.

We have investigated the first of these enough to satisfy ourselves that there is latent information on one-to-many associations in the text, and that it can be revealed by suitable extensions of our methods. However, it is clear that the combinatorial problems associated with this approach are severe, and pursuing it would require much fine tuning of the program and designing much more effective ways of indexing the most important data structures. The key to reducing the combinatorial explosion probably lies in using tables of similarities such as those the present algorithm uses to suggest combinations of items that would be worth considering. If such an approach could be made efficient enough, it is even possible that it would provide a superior way of solving the problem for which our heuristic methods of morphological analysis were introduced. Its superiority would come from the fact that it would not depend on words being formed by concatenation, but would also accommodate such phenomena as umlaut, ablaut, vowel harmony, and the nonconcatenative process of Semitic morphology.

The problems of treating recurring sequences are less severe. Data structures, such as the Patricia tree (Knuth 1973; pp. 490–493) provide efficient means of identifying all such sequences and, once identified, the data they provide could be added to

the WAT much as we now add the results of morphological analysis. Needless to say, this would only allow for uninterrupted sequences. Any attempt to deal with discontinuous sequences would doubtless also involve great combinatorial problems.

These avenues for further development are intriguing and would surely lead to interesting results. But it is unlikely that they would lead to much better sets of associations among sentences than are to be found in the SATs that our present program produces, and it was mainly these results that we were interested in from the outset. The other avenues we have mentioned concern improvements in the WAT which, for us, was always a secondary interest.

References

Baayen, H. (1991). "A stochastic process for word frequency distributions." In *Proceedings, 29th Annual Meeting of the Association for Computational Linguistics,* Berkeley, CA.

Brown, P.; Lai, J. C.; and Mercer, R. L. (1991). "Aligning sentences in parallel corpora." In *Proceedings, 29th Annual Meeting of the Association for Computational Linguistics,* Berkeley, CA.

Brown, P.; Cocke, J.; Della Pietra, S.; Della Pietra, V.; Jelinek, F.; Lafferty, J.; Mercer, R.; and Roossin P. (1990). "A statistical approach to machine translation." *Computational Linguistics,* **16**, 79–85.

Church, K. W., and Hanks, P. (1990). "Word association norms, mutual information, and lexicography." *Computational Linguistics,* **16**(1), 22–29.

Drela, M., and Langford, J. S. (1985). "Human-powered flight." *Scientific American,* **253**(5).

Drela, M., and Langford, J. S. (1986). "Fliegen mit Muskelkraft." *Spektrum der Wissenschaft.*

Fano, R. (1961). *Transmission of Information. A Statistical Theory of Communications.* MIT Press.

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

Kay, M., and Röscheisen, M. (1988). "Text-translation alignment." Technical Report, Xerox Palo Alto Research Center.

Knuth, D. E. (1973). *The Art of Computer Programming.* Vol. 3, Sorting and Searching. Addison-Wesley.

MacKeown, P. K., and Weekes, T. C. (1985). "Cosmic rays from Cygnus X-3." *Scientific American,* **253**(5).

MacKeown, P. K., and Weekes, T. C. (1986). "Kosmische Strahlen von Cygnus X-3." *Spektrum der Wissenschaft.*

van Rijsbergen, C. J. (1979). *Information Retrieval.* Butterworths.

Sato, S., and Nagao, M. (1990). "Toward memory-based translation." In *Proceedings, 15th International Conference on Computational Linguistics (COLING-90).* Helsinki, Finland.