In Proceedings of ACL Conference on Applied Natural Language Processing, 1994.

# *Termight*: Identifying and Translating Technical Terminology

**Ido Dagan** **Ken Church**
AT&T Bell Laboratories
600 Mountain Ave.
Murray Hill, NJ 07974, USA
kwc@research.att.com
dagan@research.att.com


## Abstract

Robust natural language processing algorithms, such as part of speech tagging and bilingual word alignment, are available at the research community. We suggest that both of these technologies could have an important role in glossary construction, an essential process in technical translation. Common practice among translators is to construct glossaries manually, reading hardcopies of source and target documents. This process, which is very time consuming and is usually performed under severe time constraints, yields in most cases incomplete glossaries.

We propose a semi-automatic tool, *termight*, that supports both the identification and translation of technical terms. The tool relies on part of speech tagging and word alignment programs to extract candidate terms and translations. Although the output of these programs is far from perfect it is possible to extract from it useful information that is later corrected and augmented by a user. Our extraction algorithms emphasize completeness, and identify also infrequent candidate terms and translations that may not meet some of the statistical significance criteria proposed in the literature. To make the construction process efficient we address all its aspects including procuctive interfaces that support the user's role. In particular, the tool appropriately sorts the extracted candidates, presents them to the user along with reference concordance lines, and enables fast post-processing of the candidate lists. *Termight* is currently being used by the translators at AT&T Business Translation Services.


## 1 Terminology: An Application for Natural Language Technology

The statistical corpus-based renaissance in computational linguistics has produced a number of interesting technologies such as part of speech tagging and word alignment. Unfortunately, these technologies are still not as widely used in practical applications as they might be. Part of speech taggers are used in rather few applications, such as speech synthesis (Sproat et al., 1992) and question answering (Kupiec, 1993b). Word alignment is a newer technology, available at relatively few places (Gale and Church, 1991a; Brown et al., 1993; Dagan et al., 1993). It is used at IBM for estimating parameters of their statistical machine translation prototype (Brown et al., 1993). We suggest that both of these technologies, part of speech tagging and word alignment, could have an important role in glossary construction for translation.

Glossaries are extremely important for translation. How would Microsoft, or some other software vendor, want the term "Character menu" to be translated in their manuals? Technical terms are difficult for translators because they are generally not as familiar with the subject domain as either the author of the source text or the reader of the target text. In many cases, there may be a number of acceptable translations, but it is important for consistency sake to standardize on a single one. It would be unacceptable for a manual to use a variety of synonyms to refer to the name of a menu or a button. It has become standard practice for translation houses to construct extensive job-specific glossaries to ensure consistency and correctness of technical terminology for a large job that will be translated by a number of different translators.

A glossary is a list of terms and their translations.[1] We will subdivide the task of constructing a glossary into two subtasks: (1) generating a list of terms, and (2) finding the translation equivalents. The first task

[1]The source and target fields are standard, though many other fields can also be found, e.g., usage notes, part of speech constraints, comments, etc.

will be referred to as the monolingual task and the second as the bilingual task.

How should a glossary be constructed? Translation schools teach their students to read as much background material as possible in both the source and target languages, an extremely time consuming process. The introduction to Hann (1992), a two volume text on technical translation, might help give a sense of just how laborious the standard practice is:

> Contrary to popular opinion, the job of a technical translator has little in common with other linguistic professions, such as literature translation, foreign correspondence or interpreting. Apart from an expert knowledge of both languages..., all that is required for the latter professions is a few general dictionaries, whereas a technical translator needs a whole library of specialized dictionaries, encyclopedias and technical literature in both languages; he is more concerned with the exact meanings of terms than with stylistic considerations and his profession requires certain 'detective' skills as well as linguistic and literary ones. Beginners in this profession have an especially hard time... This book attempts to meet this requirement.

Unfortunately, the academic prescriptions are often too expensive for commercial practice. Translators need just-in-time glossaries. They cannot afford to do a lot of background reading and "detective" work when they are being paid by the word. They need something more practical.

We propose a tool, *termight*, that automates some of the more tedious and laborious aspects of terminology research. The tool relies on part of speech tagging and word alignment technologies to extract candidate terms and translations. It then sorts the extracted candidates and presents them to the user along with reference concordance lines, supporting efficient construction of glossaries. The tool is currently being used by the translators at AT&T Business Translation Services (formerly AT&T Language Line Services).

*Termight* may prove useful in contexts other than human-based translation. Primarily, it can support customization of machine translation (MT) lexicons to a new domain. In fact, the arguments for constructing a job-specific glossary for human-based translation may equally hold for an MT-based process, emphasizing the need of a productivity tool. The monolingual component of *termight* can be used to construct terminology lists in other applications, such as technical writing, book indexing, hypertext linking, natural language interfaces, text categorization and indexing in digital libraries and information retrieval (Salton, 1988; Cherry, 1990; Harding, 1982; Bourigault, 1992; Damerau, 1993), while the bilingual component can be useful for information retrieval in multilingual text collections (Landauer and Littman, 1990).

## 2 Monolingual Task: An Application for Part of Speech Tagging

Although part of speech taggers have been around for a while, there are relatively few practical applications of this technology. The monolingual task appears to be an excellent candidate. As has been noticed elsewhere (Bourigault, 1992; Justeson and Katz, 1993), candidate technical terms can be found by looking for repeated multiword noun phrases. An expert terminologist can then skim the list of candidates to weed out spurious candidates and cliches.

Very simple procedures of this kind have been remarkably successful. They can save an enormous amount of time over the current practice of reading a hardcopy of the document to be translated, focusing on tables, screen dumps, figures, index, table of contents and so on, and writing down terms that happen to catch the translator's eye. The current process is not only very laborious, but also runs the risk of missing large numbers of important terms.

We follow the spirit of these efforts. A part of speech tagger (Church, 1988) is used to identify a list of candidate terms which is then filtered by a manual pass. We have found, however, that the manual pass dominates the cost of the monolingual task, and consequently, we have tried to design an interactive user interface (see Figure 1) that minimizes the burden on the expert terminologist. The terminologist is presented with a list of candidate terms, and corrects the list with a minimum number of key strokes. The interface is designed to make it easy for the expert to pull up relevant evidence to help identify incorrect candidates as well as terms that are missing from the list. The candidates are sorted so that the better ones are found near the top of the list, and so that related candidates appear near one another.

### 2.1 Candidate terms and associated concordance lines

**Candidate terms.** The list of candidate terms contains both multi-word noun phrases and single words. The multi-word terms match a small set of syntactic patterns defined by regular expressions and are found by searching a version of the document tagged with parts of speech (Church, 1988). Single word candidates are defined by taking the list of all words that occur in the document and do not appear in a standard stop-list of "noise" words.

**Grouping and sorting of terms.** The list of candidate terms is sorted to group together all noun phrase terms that have the same head word (as in Figure 1).[2] The order of the groups in the list is de-

[2] We currently prefer to include multi-word noun


Figure 1: The monolingual user interface consists of three screens: (1) the list of candidate terms (upper right), (2) the final list of terms, as constructed by the user (upper left), and (3) the concordance lines associated with the current term. Typos are due to OCR errors. Underscores denote line breaks.

termined by decreasing frequency of the head word in the document, which usually correlates with the likelihood that this head word is used in technical terms.

**Sorting within groups.** Under each head word the terms are sorted alphabetically according to reversed order of the words. Sorting in this order reflects the order of modification in simple English noun phrases and groups together terms that denote different modifications of a more general term (see for example the terms *default paper size, paper size* and *size* in Figure 1).

**Concordance lines.** To decide whether a candidate term is indeed a term, and to identify multi-word terms that are missing from the candidate list, one must view relevant lines of the document. For this purpose we present a concordance line for each occurrence of a term (a text line centered around the term). If, however, a term, $t_1$, (like 'point') is contained in a longer term, $t_2$, (like 'insertion point' or 'decimal point') then occurrences of $t_2$ are not displayed for $t_1$. This way, the occurrences of a general term (or a head word) are classified into disjoint sets corresponding to more specific terms, leaving only unclassified occurrences under the general term. In the case of 'point', for example, five specific terms are identified that account for 61 occurrences of 'point', and accordingly, for 61 concordance lines. Only 20 concordance lines are displayed for the word 'point' itself, and it is easy to identify in them 5 occurrences of the term 'starting point', which is missing from the candidate list (because 'starting' is tagged as a verb). To facilitate scanning the concordance lines they are sorted such that all occurrences of identical preceding contexts of the head word, like 'starting', are grouped together. Since all the words of the document, except for stop list words, appear in the candidate list as single word terms it is guaranteed that every term that was missed by the automatic procedure will appear in the concordance lines.

phrases that consist only of a sequence of nouns, as these seem to yield the best hit rate in our environment. In such noun phrases, the head word is always the last noun.

## 2.2 User interface

The monolingual user interface is divided to three screens. The upper-right screen displays the list of candidate terms. The bottom screen displays the concordance lines of the current candidate term, as indicated by the cursor position in the first screen. The upper-left screen is used for the final term list, as constructed by the user. A single key-press copies the current candidate term, or the content of any marked emacs region, into the upper-left screen.

## 2.3 Evaluation

Using the monolingual component, a terminologist at AT&T Business Translation Services constructs terminology lists, as the first step of a translation job, at the impressive rate of 150-200 terms per hour. For example, It took about 10 hours to construct a list of 1700 terms extracted from a 300,000 word document. The deployment of the tool has at least doubled the rate of constructing terminology lists, which was previously performed by simpler lexicographic tools.

### 2.4 Comparison with related work

We have found that addressing the human's role in our system has influenced the design of the machine's role. In particular, we decided to sort the terms by decreasing frequency of their head word rather than using mutual information (MI) scores or similar statistics which dominate recent literature (Church and Hanks, 1990; Smadja, 1993; Wu and Su, 1993).[3] First, frequent head words are likely to provide many frequent terms, and are therefore more important for the glossary (a "productivity" criterion). Second, MI-style statistics measure the salience of a term relative to the overall frequency of its component words, assigning high scores only to terms that are relatively frequent compared with other terms with which they share a certain word. This effect is magnified for frequent words, such as *box* in the Microsoft Windows manual, resulting in very low MI scores for infrequent terms that contain this word. We have found, however, that most candidate terms with a frequent head word, including infrequent terms, are correct.[4] For example, the terms 'Format Disk dialog box' and 'Label Disk dialog box' occur only once in the document while it is clear that they should be part of its terminology list. With our sorting, these terms appear at the beginning of the list with all other terms of the head word 'box'.

Another characteristic of previous work is that each candidate term is scored independently of other terms. We need to score a group of related terms rather than each term at a time. Future work may enhance our simple head word frequency score and may take into account additional relationships between terms, including common words in modifying positions.

[3]An exception is found in (Justeson and Katz, 1993), where the candidate list is sorted by decreasing *term* frequency (compared with our sorting by decreasing *head word* frequency).

[4]It should be noted that our observation relates to identifying domain specific terminology, while some of the work on MI-style statistics was proposed for large, domain independent, corpora.

## 3 Bilingual Task: An Application for Word Alignment

### 3.1 Sentence and word alignment

Aligning parallel texts has recently received considerable attention (Warwick et al., 1990; Brown et al., 1991a; Brown et al., 1993; Gale and Church, 1991b; Gale and Church, 1991a; Kay and Roscheisen, 1993; Simard et al., 1992; Church, 1993; Kupiec, 1993a; Matsumoto et al., 1993; Dagan et al., 1993). These methods have been used in statistical machine translation (Brown et al., 1990), terminology research and translation aids (Isabelle, 1992; Ogden and Gonzales, 1993; van der Eijk, 1993), bilingual lexicography (Klavans and Tzoukermann, 1990; Smadja, 1992), word-sense disambiguation (Brown et al., 1991b; Gale et al., 1992) and information retrieval in a multilingual environment (Landauer and Littman, 1990).

Most of the work on developing alignment algorithms, as well as on applying them, was concerned with alignment at the sentence level. Algorithms for the more difficult task of word level alignment were proposed in (Gale and Church, 1991a; Brown et al., 1993; Dagan et al., 1993) and were applied at IBM for estimating parameters of their statistical machine translation prototype (Brown et al., 1993).

In terminology research it was recognized that previously translated texts provide a major source of information about technical terms. As argued by Isabelle (1992), "Existing translations contain more solutions to more translation problems than any other existing resource." Even if other resources are available, such as general technical dictionaries, it is important to verify the translation of terms in previously translated documents of the same customer (or domain) to ensure consistency across documents.

Several translation workstations provide sentence alignment and allow the user to search interactively for term translation in aligned archives (e.g. (Ogden and Gonzales, 1993)). It was also suggested to partially automate the acquisition process (Picchi et al., 1992). A couple of methods propose to use sentence alignment and additional statistics to find candidate translations of terms (Smadja, 1992; van der Eijk, 1993).

We suggest that word level alignment may be better suitable for the term translation task. The bilingual component of *termight* gets as input a list of source terms and a bilingual corpus aligned at the word level. We have been using the output of *word_align*, a robust alignment program that proved useful for bilingual concordancing of noisy texts (Dagan et al., 1993). *Word_align* produces a partial mapping between the words of the two texts, skipping words that cannot be aligned at a given confidence level (see Figure 2).

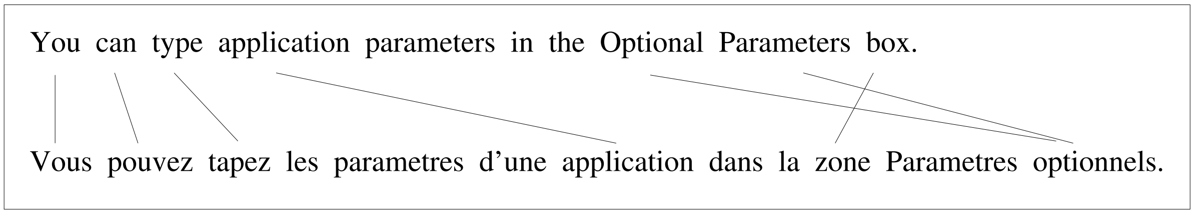


Figure 2: An example of *word_align*'s output for the English and French versions of the Microsoft Windows manual. The alignment of *Parameters* to *optionnels* is an error.

### 3.2 Candidate translations and associated concordance lines

For each occurrence of a source term *termight* identifies a *candidate translation* based on the alignment of its words. The candidate translation is defined as the sequence of words that spans between the first and last target positions that are aligned with any of the words of the source term. In the example of Figure 2 the candidate translation of *Optional Parameters box* is *zone Parametres optionnels*, since *zone* and *optionnels* are the first and last French words that are aligned with the words of the English term. Notice that in this case the candidate translation is correct even though the word *Parameters* is aligned incorrectly. In other cases alignment errors may lead to an incorrect candidate translation for a specific occurrence of the term. It is quite likely, however, that the correct translation, or at least a string that overlaps with it, will be identified in some occurrences of the term.

*Termight* collects the candidate translations from all occurrences of a source term and sorts them in decreasing frequency order. The sorted list is presented to the user, followed by bilingual concordances for all occurrences of each candidate translation (see Figure 3). The user views the concordances to verify correct candidates or to find translations that are missing from the candidate list. The latter task becomes especially easy when a candidate overlaps with the correct translation, as it directs the attention of the user to the concordance lines of this particular candidate, which are likely to be aligned correctly. A single key-stroke copies a verified candidate, or a translation that is marked in a concordance line, into the appropriate place in the glossary.

### 3.3 Evaluation

An evaluation of the user-interface found that the correct answer was often the first choice (40%) or the second choice (7%) in the candidate list. For the remaining 53% of the terms, the correct answer was always somewhere in the concordances. Using the interface, the glossary was translated at a rate of about 100 terms per hour based on the terms found in the English and German versions of a technical manual.

### 3.4 Related and future work

Smadja (1992) and van der Eijk (1993) describe term translation methods that use bilingual texts that were aligned at the sentence level. Their methods find likely translations by computing statistics on term cooccurrence within aligned sentences and selecting source-target pairs with statistically significant associations. We found it advantageous to rely on explicit word alignments. It enables us to identify translations of infrequent terms that would not otherwise meet statistical significance criteria. If the words of a term occur at least several times in the document (regardless of the *term* frequency) then *word_align* is likely to align them correctly and *termight* will identify the correct translation. If only some of the words of a term are frequent then *termight* is likely to identify a translation that overlaps with the correct one, directing the user quickly to correctly aligned concordance lines. Even if all the words of the term were not aligned by *word_align* it is still likely that most concordance lines are aligned correctly based on other words in the near context.

*Termight* motivates future improvements in word alignment quality which will increase recall and precision of the candidate lists. In particular, alignment algorithms that take into account syntactic structure will impose more restrictions on possible alignments of complete terms.

Finally, *termight* can be extended for verifying translation consistency at the proofreading (editing) step of a translation job, after the document has been translated. For example, in an English-German document pair the tool identified the translation of the term *Controls menu* as *Menu Steuerung* in 4 out of 5 occurrences. In the fifth occurrence *word_align* failed to align the term correctly because another translation, *Steuermenu*, was uniquely used, violating the consistency requirement. *Termight*, or a similar tool, can thus be helpful in identifying inconsistent translations.

## 4 Conclusions

We have shown that terminology research provides a good application for robust natural language technology, in particular for part of speech tagging and word alignment algorithms. Although the output of

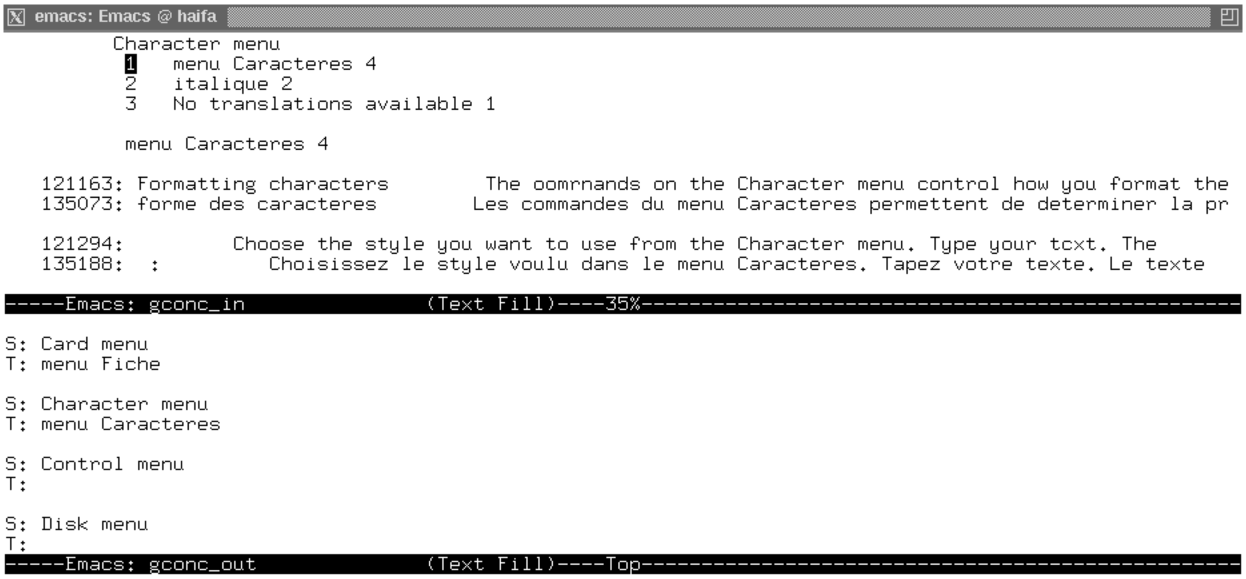


Figure 3: The bilingual user interface consists of two screens. The lower screen contains the constructed glossary. The upper screen presents the candidate translations of the current term, followed by bilingual concordances for each candidate (the number following a candidate denotes its frequency). Typos are due to OCR errors.

these algorithms is far from perfect it is possible to extract from it useful information that is later corrected and augmented by a user. Our extraction algorithms emphasize completeness, and identify also infrequent candidates that may not meet some of the statistical significance criteria proposed in the literature. To make the process efficient, however, it is necessary to address all its aspects, including appropriate interfaces that support the user's role. In many cases, improving the way information is presented to the user may have a larger effect on productivity than improvements in the underlying natural language technology. In particular, we have found the following to be very effective:

- Grouping linguistically related terms, making it easier to judge their validity.
- Sorting candidates such that the better ones are found near the top of the list. With this sorting one's time is efficiently spent by simply going down the list as far as time limitations permit.
- Providing quick access to relevant concordance lines to help identify incorrect candidates as well as terms or translations that are missing from the candidate list.

As the need in efficient knowledge acquisition tools becomes widely recognized, we hope that this experience with *termight* will be found useful for other text-related systems as well.

## References

Didier Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proc. of COLING*, pages 977–981.

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to language translation. *Computational Linguistics*, 16(2):79–85.

P. Brown, J. Lai, and R. Mercer. 1991a. Aligning sentences in parallel corpora. In *Proc. of the Annual Meeting of the ACL.*

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1991b. Word sense disambiguation using statistical methods. In *Proc. of the Annual Meeting of the ACL*, pages 264–270.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.

L. L. Cherry. 1990. Index. In *Unix Research System Papers*, volume 2, pages 609–610. AT&T, 10 edition.

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

Kenneth W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text.

In *Proc. of ACL Conference on Applied Natural Language Processing.*

Kenneth W. Church. 1993. Char_align: A program for aligning parallel texts at character level. In *Proc. of the Annual Meeting of the ACL.*

Ido Dagan, Kenneth Church, and William Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1–8.

Fred J. Damerau. 1993. Generating and evaluating domain-oriented multi-word terms from texts. *Information Processing & Management*, 29(4):433–447.

William Gale and Kenneth Church. 1991a. Identifying word correspondence in parallel text. In *Proc. of the DARPA Workshop on Speech and Natural Language.*

William Gale and Kenneth Church. 1991b. A program for aligning sentences in bilingual corpora. In *Proc. of the Annual Meeting of the ACL.*

William Gale, Kenneth Church, and David Yarowsky. 1992. Using bilingual materials to develop word sense disambiguation methods. In *Proc. of the International Conference on Theoretical and Methodolgical Issues in Machine Translation*, pages 101–112.

Michael Hann. 1992. *The Key to Technical Translation*, volume 1. John Benjamins Publishing Company.

P. Harding. 1982. Automatic indexing and classification for mechanised information retrieval. BLRDD Report No. 5723, British Library R & D Department, London, February.

P. Isabelle. 1992. Bi-textual aids for translators. In *Proc. of the Annual Conference of the UW Center for the New OED and Text Research.*

John Justeson and Slava Katz. 1993. Technical terminology: some linguistic properties and an algorithm for identification in text. Technical Report RC 18906, IBM Research Division.

M. Kay and M. Roscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1):121–142.

J. Klavans and E. Tzoukermann. 1990. The bicord system. In *Proc. of COLING.*

Julian Kupiec. 1993a. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proc. of the Annual Meeting of the ACL.*

Julian Kupiec. 1993b. MURAX: A robust linguistic approach for question answering using an on-line encyclopedia. In *Proc. of International Conference on Research and Development in Information Retrieval, SIGIR*, pages 181–190.

Thomas K. Landauer and Michael L. Littman. 1990. Fully automatic cross-language document retrieval using latent semantic indexing. In *Proc. of the Annual Conference of the UW Center for the New OED and Text Research.*

Yuji Matsumoto, Hiroyuki Ishimoto, Takehito Utsuro, and Makoto Nagao. 1993. Structural matching of parallel texts. In *Proc. of the Annual Meeting of the ACL.*

William Ogden and Margarita Gonzales. 1993. Norm – a system for translators. Demonstration at ARPA Workshop on Human Language Technology.

Eugenio Picchi, Carol Peters, and Elisabetta Marina. 1992. A translator's workstation. In *Proc. of COLING*, pages 972–976.

Gerard Salton. 1988. Syntactic approaches to automatic book indexing. In *Proc. of the Annual Meeting of the ACL*, pages 204–210.

M. Simard, G. Foster, and P. Isabelle. 1992. Using cognates to align sentences in bilingual corpora. In *Proc. of the International Conference on Theoretical and Methodolgical Issues in Machine Translation.*

Frank Smadja. 1992. How to compile a bilingual collocational lexicon automatically. In *AAAI Workshop on Statistically-based Natural Language Processing Techniques*, July.

Frank Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177.

R. Sproat, J. Hirschberg, and D. Yarowsky. 1992. A corpus-based synthesizer. In *Proceedings of the International Conference on Spoken Language Processing*, pages 563–566, Banff, October. ICSLP.

Pim van der Eijk. 1993. Automating the acquisition of bilingual terminology. In *EACL*, pages 113–119.

S. Warwick, J. Hajic, and G. Russell. 1990. Searching on tagged corpora: linguistically motivated concordance analysis. In *Proc. of the Annual Conference of the UW Center for the New OED and Text Research.*

Ming-Wen Wu and Keh-Yih Su. 1993. Corpus-based compound extraction with mutual information and relative frequency count. In *Proceedings of ROCLING VI*, pages 207–216.