

Using Word Alignment to Extend Multilingual Medical Terminologies

Louise Deléger*, Magnus Merkel†, Pierre Zweigenbaum*‡

*Inserm U729, Paris, France

†Dept of Computer and Information Science, Linköping University, Sweden

‡Assistance Publique – Paris Hospitals, STIM; Inalco, CRIM, Paris, France

louise.deleger@spim.jussieu.fr, mme@ida.liu.se, pz@biomath.jussieu.fr

Abstract

Medical terminologies such as those provided in the UMLS are never exhaustive and there is a constant need to enrich them, especially in terms of multilinguality. We present a methodology to acquire new French translations of English medical terms based on word alignment in a parallel corpus — i.e. pairing of corresponding words. We automatically collected a 27.7-million-word parallel, English-French corpus. Based on a first 1.3-million-word extract of this corpus, we detected 3,255 French translations of English MeSH terms, among which 1,956 are new translations.

1. Introduction

The UMLS Metathesaurus is an extensive vocabulary database that gathers and provides a link between different existing biomedical terminologies. But despite being a multilingual resource, it is mostly composed of English vocabulary, and other languages such as French are under-represented in comparison to English. There is therefore a need to enrich the terminologies of the UMLS. The acquisition of new translations of English terms is required. This is the purpose of the VUMeF¹ project which aims at extending the French part in the UMLS and which provides the background for this work.

Plenty of multilingual texts can be found as regards a specific domain but exhaustive terminologies and dictionaries are far less numerous — as can be seen in the case of the UMLS. Hence the idea of using parallel corpora (collections of multilingual texts) to enlarge terminologies. So instead of employing a human translator, we can make use of existing translated texts from which translations at the term level can be extracted.

We present a methodology to acquire medical terms based on word alignment in a parallel corpus. Word alignment is a natural language processing technique and is used in several applications such as terminology development (which is the case here), automatic translation and cross-lingual information retrieval. It consists in pairing words that are translations of each other in a parallel corpus.

Previous work has addressed the issue of multilingual medical terminologies. Chiao and Zweigenbaum (2002) collect translations from comparable corpora. Baud et al. (1998) make use of already parallel medical vocabularies to derive word translations. Widdows et al. (2002) use a statistical vector model on a corpus aligned at the document level. Névéal and Ozdowska (2005) have an approach similar to ours in that it deals with word alignment in parallel medical corpora to extract French translations of English terms. However, we deal with a larger corpus and process all kinds of alignments.

1. Project VUMeF (French Unified Medical Vocabulary), led by Stéfan Darmoni (Darmoni et al., 2003), is partially funded by the French Ministry of Research (National Network of Health Technologies).

Our task involves issues such as dealing with errors in the alignment process that will spread from step to step, and detecting multi-word units — a term being either a single word or a multi-word expression.

This work is outlined in the following way: based on a French-English corpus (2.1), we align sentences (2.2) and words (2.3, 2.4). Medical terms are then selected (2.5) through the projection of a list of English terms from the MeSH. We obtain a list of bilingual English-French medical terms that we review. We extract samples for evaluation purposes (2.6) and expose results (3.). We discuss (4.) and conclude (5.) on the method.

2. Material and Method

2.1. Corpus

The corpus used for this experiment is collected from the web. The web is indeed a powerful resource for building corpora, both in terms of quantity and multilinguality. The quality of such a corpus can nevertheless be questioned and this might account for a proportion of the noise detected in the results.

Our corpus is gathered from a bilingual (French-English) Canadian health web site². It is intended for the general public as well as for specialists and the proportion of specialized terms might therefore be lower than in resources dedicated to medical specialists.

Several techniques exist for building a parallel corpus from the web (Resnik and Smith, 2003; Patry and Langlais, 2005). We generate pairs of parallel documents (i.e., documents that are translations of each other) using information contained in the document structure — namely, HTML links to corresponding documents in the other language. Indeed, after a study of the documents, we noticed that each document provided access to its translation page through an image or a text tag labelled in the corresponding language (specified in the “alt” attribute of the HTML tag). This gives us 11,041 pairs of parallel documents and a total of 27.7 million words.

Documents from the web usually come in either HTML or PDF format, and need to be converted to text format. As

2. <http://www.hc-sc.gc.ca/>

for us, we have HTML documents and thus have access to structural information that may be useful to keep even in a text file. After cleaning the HTML files and converting them to XML format, we use a XSLT stylesheet to transform them to text format while keeping a number of information — title, paragraph and link tags which will be used as correspondence points for sentence alignment. Indeed, we assume that a title in a source language corresponds to a title in a target language, a link to a link, and in most cases a paragraph to a paragraph. The resulting texts are segmented into sentences to prepare the way for further processing.

2.2. Sentence Alignment

The first step towards word alignment is to align the corpus at sentence-level. Sentence alignment is a mandatory task since there is not a full one-to-one correspondence between the sentences of two parallel texts. Although it is most common that one sentence in a source language corresponds to one sentence in a target language, there are instances where one sentence is translated with two — or sometimes even three or more — sentences, and this needs to be determined before working at the word level.

To do so, we use Dan Melamed's GMA³ (Geometric Mapping and Alignment) (Melamed, 2000), a robust tool which performs sentence alignment of parallel texts using both statistical and linguistic techniques. It is based, among other things, on length measurements, bilingual lexicons and cognates (words sharing similar spelling and meaning). Though sentence aligners in general and GMA in particular achieve high-quality performances, any mistake at this level will be reflected at the next one — i.e., word alignment — and will make things even harder for this already complex process. So, in order to work on cleaner data, we attempt to automatically detect and remove incorrect sentence alignments as well as bad document pairing (documents that are not parallel) using criteria such as sentence length and quality evaluation of sentence alignment.

2.3. Word Alignment

Once sentences are aligned, we can proceed to word alignment. This task is far more problematic than sentence alignment. There is no true word-to-word correspondence between the words of two sentences. A word is often translated with several ones, or can be omitted in the corresponding sentence (this is typically the case for grammatical words that are specific to a language). Parallel sentences, though being translations of each other, can differ considerably in terms of structure. In that case even a human has trouble determining which words should be paired together. The results we expect are therefore on a lower level than from the previous sentence alignment task.

The issue of the type of word alignment should be raised. That is, are we satisfied with a word-to-word alignment? The objective of this work is to obtain medical terms. A term can be either a single word or a multi-word unit. A common approach is to first extract candidate terms using a separate tool — a term extractor — and then to proceed to their alignment (Daille et al., 1994; Gaussier, 1998). The

originality of this work is that we do not separate the detection of candidate terms from the alignment process. In other words we use a tool that is able to detect multi-word units and to align both single words and multi-word expressions.

Word alignment systems usually derive from either statistical approaches or linguistic ones, or a combination of both. Statistical methods (Brown et al., 1993) involve co-occurrence measures and probability scores, and are especially effective on large corpora with high-frequency words but performances decrease with low-frequency occurrences. Linguistic ones (Wu, 2000) make use of information such as syntactic parsing. They are less robust despite being able to deal with low-frequency words. Hybrid approaches (Ahrenberg et al., 2000; Barbu, 2004) seem to be a good compromise.

2.4. Aligning Words with the I*Tools

We use the I*Tools suite (developed at Linköping University, Sweden) to perform word alignment. We chose these tools partly because they are based on a hybrid approach, using both statistical and linguistic techniques. They also align multi-word units which suits our terminological purpose. They make use of resources such as co-occurrence measures, bilingual dictionaries, POS tagging and syntactical analysis.

A pre-processing step is required: the corpus is tagged and lemmatized (using Treetagger⁴) and syntactically annotated (with the syntactic analyzer SYNTAX (Bourigault et al., 2005)). The files are transformed into XML format encoding this information.

The alignment process with the I*Tools can be divided into three steps: training, automatic alignment and review of the results; each one corresponding to a specific tool of the suite. Training and review are both done with graphical, interactive tools that are fast to work with.

Training of the system is manually done using a special tool of the suite — the I*Link⁵ interactive aligner (Merkel et al., 2003). This tool proposes word pairings to the user who accepts or rejects them. The user's decisions are stored into the resources of the system and by learning from them, the performances become increasingly accurate. These resources provide training data for the automatic word aligner.

The corpus is then automatically aligned by I*TriX, the automatic aligner of the suite, using the resources created with I*Link. We obtain a list of word alignments — i.e., source words paired with target words. The system can also exploit data created during the next step (the reviewing phase). In that case, the automatic alignment is repeated after a first run and takes into account the review made by the user. This is useful if the results first obtained are not as good as expected.

Results are reviewed with the I*View tool which enables the user to confirm, reject, or simply remove an alignment. An alignment is « removed » when it is neither an error nor a correct alignment, meaning it is a partial alignment (some

3. <http://nlp.cs.nyu.edu/GMA/>

4. <http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html>

5. <http://www.ida.liu.se/~nplab/ILink/>

parts are correct). This tool also indicates for each alignment a quality score, which enables the user to rank the alignments. The quality score used in the I*Tools is based on the mutual information formula (Stolz, 1965). Mutual information has been used in several works, including (Church and Hanks, 1990) which derives a new measure for estimating word associations and (Fabry et al., 2005) which uses mutual information for term extraction to build a terminology. In our case, the measure is expressed in terms of the frequency of the words as a pair, and the frequency of each source and target word of the pair independently in other pairs. This means that for a proposed word pair which occurs with a high frequency and where the source word and the target word only occur in this pair and not in any other suggested word pairs, we have good reasons to assume that the quality is high. On the other hand if the word pair has a low frequency and the source and target words of the pair are found in several other suggested pairs, then there is reason to be more doubtful to the suggestion. The formula is $Q=f(st)/n(s)+n(t)$, where $f(st)$ is the frequency of the word pair and $n(s)$ and $n(t)$ are the number of different word pairs in which the source and target words occur respectively.

2.5. Term Selection and review

In practice, there is no need to review all of the results since we are only looking for medical terms. We thus retain only those likely to be of interest. We select them using an English medical terminology — namely MeSH (as extracted from the 2005AC version of the UMLS). We project this list of terms onto the English entries of our alignment pairs and select those present in the pairs. Only then do we review the alignments. These alignments constitute French candidate translations of English MeSH terms. The review can be done by a linguist engineer. Afterwards, we can determine the proportion of new translations retrieved and submit these translations to medical experts.

In order to restrain manual interaction, we also tested a different solution, that is, not reviewing the results directly after the term selection, but only after some filtering — elimination of duplicates, verbs, translations already existing in the French version of the MeSH, and alignments with a poor quality score. Indeed, we consider that MeSH terms are mainly nouns and that verbs are not needed in the selection — they introduce too much noise. Low quality score alignments are also likely to be errors and might not be worth reviewing. This filtering phase will reduce the amount of terms to be reviewed.

2.6. Implementation and Evaluation

The methodology described above was implemented as follows:

1. conversion of the corpus into text format;
2. sentence alignment;
3. training of the automatic word aligner on a set of 600 sentences randomly taken from our corpus, by interacting manually with I*Link;
4. automatic word alignment with I*Trix. If results seem poor, a first review may also be done followed by a second run of I*Trix.

5. selection of medical terms (see section 2.5);

Implementation 1:

6. review, with I*View, of the alignments for the terms selected.

Implementation 2:

6. filtering of the results:
 - elimination of duplicates
 - elimination of verbs (the MESH entries are considered to be nouns)
 - elimination of terms already registered in the French MeSH
 - elimination of pairs with a poor quality score (equal to 0); however, there may be correct alignments among the low quality score ones. In that case, a small proportion of translations will be lost. We have tested the implementation both with this step and without it.
7. review of the results.

Evaluation was performed at several points of the implementation. First, we performed an evaluation of the quality of the alignment at step 2 by checking 100 sentences randomly taken from the corpus and measuring the percentage of correct alignments (precision measure).

The quality of word alignment was evaluated at step 4 by measuring precision on two samples: sample 1 consists of 100 word pairs randomly taken from the whole resulting pairing, and sample 2 of 100 word pairs taken from the best word alignments (alignments with a frequency higher than 1 and a good quality score — equal to or higher than 1).

Last, step 6 of implementation 1 allowed us to have a gold standard to evaluate word alignment for the medical terms. We evaluated the performances using information retrieval evaluation techniques, namely precision-recall measures. Other teams have also used these measures for evaluating tasks aside from information retrieval — text categorization for instance (Larkey and Croft, 1996). In information retrieval, precision is computed at 11 recall points from a list of retrieved documents. In our case, we used a ranked list of alignments instead of documents, considering that an alignment being correct is similar to a document being relevant for a query. We used `trec_eval`⁶ to compute these recall points and obtained a precision-recall curve. These measures are calculated on the basis of the alignments ranked by frequency and quality score, meaning that the first alignments are expected to be the best ones. These recall-precision points also allowed us to measure the mean average precision.

This step was also useful to determine the proportion of errors and correct alignments in the filtered results at step 6 of implementation 2, thereby allowing us to experiment with the setting of a threshold for the quality score of the alignments to be filtered out.

3. Results

3.1. Valid Alignments (Language Engineer)

We completed steps 1 and 2 on the whole corpus, thus obtaining 1.1 million sentence pairs. As the corpus is huge,

6. http://trec.nist.gov/trec_eval/trec_eval.8.0.tar.gz

	Sample 1	Sample 2	Set of medical terms
Precision	50%	92.2%	52%
Errors	19.6%	4.9%	30.3%
Partial alignments	30.4%	2.9%	17.7%

Table 1: Evaluation figures for word alignment

we have currently processed only part of it from step 3 to the end — a set of 540 pairs of documents (1.3 million words) gathered in two corresponding files. From this set, we obtained 91,171 word alignments and selected 10,392 pairs of medical terms.

Among these pairs, there are 2,567 different source terms (a term can have several translations), so we have a mean value of 4.05 French translations per English term. 5,403 alignments were confirmed as correct ones — by a language engineer (LD) — which gives a precision of 52% (see table 1) and a mean value of 2.1 correct translations per term. We count 2,159 different source terms in the confirmed alignments, meaning that 408 terms only had incorrect translations.

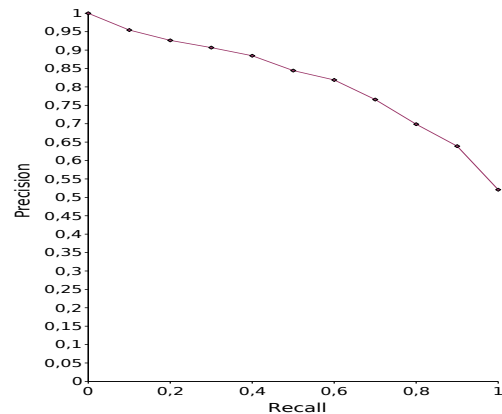
Table 1 shows that evaluation results for the overall quality of word alignments (step 4 of 2.6) are very good for the top alignments (sample 2, taken from the 7,366 best alignments as described in 2.6) and average for the whole aligned corpus (sample 1). As for sentence alignment (step 2 of 2.6), we achieved a precision of 95%, which is excellent.

Precision-recall figures for the evaluation of the set of medical terms (as described in 2.6, step 6) with *trec_eval* are detailed in the table on figure 1 and emphasize the previous statement that precision is excellent for the first alignments and decrease afterwards. To be more accurate, these recall points were computed on a scale of 10,000 alignments instead of the standard of 1,000 documents used in information retrieval. The increasingly descending slope of the curve on figure 2 shows that the ranking algorithm does push the majority of incorrect alignments towards the end of the list, with an inflection around 60% recall, obtaining more than 80% precision. The mean average precision measured is indeed 82%.

A proportion of the noise in word alignment can be attributed to errors in the sentence alignment process: 17% of the incorrect alignments are due to bad sentence pairing. Other factors include errors in POS tagging, bad document pairing (in our case we observed some English-English document pairs) and low quality of the data — misspelling of words, insertion of spaces inside a word, missing spaces between words.

3.2. Useful Medical Translations (Medical Specialist)

If we take a look at the resulting list of 5,403 confirmed medical term alignments, we notice 306 pairs that are not real translations but merely pairs of English words — i.e. the English words have not been translated. These are considered correct alignments but are of no use for our purpose, so we simply ignore them. Among the remaining 5,097, we eliminate a number of duplicates (pairs that are

Table 2: Precision at 11 recall levels, measured with *trec_eval* on a scale of 10,000 alignments

the same but were not considered as such by the alignment tools due to case differences) and obtain 3,607 word pairs. As stated in 2.5, we do not consider verbs as valid candidate translations and we eliminate them, thus lowering down the number of translations to 3,255. In this set, we look at the number of different concepts (CUI in the UMLS), terms already present in the French version of the MeSH and new translations (see table 3). The translations include morphological variants — for instance adjectives instead of prepositional phrases — and synonyms. However we do not consider plural/singular and masculine/feminine morphological variants as new translations.

A sample of 145 MeSH terms (see figures in table 3) was also extracted for validation purposes. 79 terms had new translations which were submitted to expert validation. 64 have been validated. Examples of translations are given in table 4.

	Complete set of MeSH terms	Validation sample
Translation pairs	3,255	145
Concepts	1,868	138
Already registered	1,299	66
New	1,956	79
New and valid		64 (81%)

Table 3: Figures for the MeSH translations

English	French	Valid
bone cancer	cancer des os	Yes
breast milk	lait maternel	Yes
reproduction rights	droits de reproduction	No

Table 4: Translation examples

The second implementation tested — i.e., reviewing only the filtered results — gives the following figures. From the 10,392 pairs of selected terms, we eliminate duplicates (8,699 resulting pairs), verbs (7,985 resulting pairs), and select only the new translations (not registered in the French

MeSH), which gives us 6,670 candidate translations to be reviewed. Thanks to the first implementation, we can easily determine the proportion of noise. Since we expect 1,956 new translations, 4,714 should be eliminated. Among these, there are incorrect alignments (4,452) and English-English pairs (262). We can see that the precision is very low, but that was expected. Since we are only looking at new translations pairs, incorrect word alignments are bound to be considered as new. Most of the noise is therefore selected and we do have the same balance as in implementation 1. Evaluation of the alignment is here meaningless and should only be performed on the non-filtered results. But this implementation enables us to considerably reduce the amount of word pairs to be reviewed: we have to review 6,670 alignments instead of 10,392, that is, 35.8% less.

We also tested filtering out alignments with a low quality score. We set the threshold to 0: all alignments with a score equal to 0 are removed. With this criterion, we lower down the selection to 4,493 alignments, which means that we now have 56.8% less to review and 2,766 noisy alignments. Among the 2,177 removed, 238 were actually correct. So there is a loss of information, but the proportion remains small (12%). Alignments with a score equal to 0 have an 89% chance of being incorrect, which can justify for their being removed.

4. Discussion

Our approach presents a number of advantages as well as some drawbacks. It allows us to acquire medical terms which are actually used in French documents for certain MeSH descriptors but are not registered in the current French version of the MeSH. It does not require a human translator and makes the best of existing resources. In terms of word alignment, we are able to process noisy data quite efficiently. We do not use monolingual term extractors and we align single words and multiword expressions with a uniform approach unlike other methods which concentrate on 1-1 and 2-2 word alignments (Névél and Ozdowska, 2005).

Though being an automatic approach, it still needs human help in the process (training and validation). The success of this method is also heavily dependent on the efficiency of word alignment which is a complex task. However, the remaining processing of the rest of the medical web corpus, if done incrementally, could steadily increase the quality of word alignment. Using the techniques outlined in this paper to minimize the reviewing process it should be possible to rapidly include verified data in each step and include this as positive training data for each new iteration. If the corpus is divided into roughly 25 sets containing just over a million words per set, and these subcorpora were processed one by one, with a short reviewing process included, the confirmed entries of each run could be fed into the training data for the next run of the automatic alignment. We also assume that interactive training using I*Link will not be necessary for each subsequent iteration, which means that the manual time spent on each new iteration will decrease and the precision will likely increase due to new training data.

The quality of the corpus is an important feature and its

choice is a major issue. In our case, we used the Web as a resource and processed a whole website. A study of the documents would have been useful in order to best characterize the type of data acquired — which documents are intended for medical specialists, which ones are for the general public and which ones have no medical content (index pages for instance). Interesting developments of this method will include the specific search for patient-oriented translations (consumer vocabulary) which are even more lacking in medical terminologies. This can be achieved, for instance, to look for candidate translations of Medline Plus vocabulary.

5. Conclusion

We described a methodology to acquire new translations of English medical terms in order to enrich existing medical terminologies. We argued that a natural language processing technique such as word alignment is an efficient way to do so. Indeed, we were able to find a number of new translations of English MeSH terms. Moreover, it is an automatic process which only requires limited human intervention. Finally, this method raises interesting prospects such as the acquisition of patient vocabulary, and more generally its application to other parallel corpora.

Acknowledgements

We thank Stéfan Darmoni for reviewing the candidate MeSH translations, and Didier Bourigault for providing SYNTEX.

6. References

- Lars Ahrenberg, M Andersson, and Magnus Merkel. 2000. A knowledge-lite approach to word alignment. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*. Springer.
- Ana-Maria Barbu. 2004. Simple linguistic methods for improving a word alignment algorithm. In *Proceedings 7th International Conference on the Statistical Analysis of Textual Data*, pages 88–98, Louvain-la-Neuve, Belgium. Presses universitaires de Louvain.
- Robert H Baud, C Lovis, AM Rassinoux, PA Michel, and Scherrer JR. 1998. Automatic extraction of linguistic knowledge from an international classification. In C Safran B Cesnik and P Degoulet, editors, *Proc 9th World Congress on Medical Informatics*, pages 581–5.
- Didier Bourigault, Cécile Fabre, Cécile Fréerot, Marie-Paule Jacques, and Sylwia Ozdowska. 2005. Syntex, analyseur syntaxique de corpus. In *Proceedings Traitement automatique des langues naturelles (Traitement automatique des langues naturelles)*, Dourdan.
- PF Brown, SAD Pietra, VJD Pietra, and RL Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311.
- Yun-Chuang Chiao and Pierre Zweigenbaum. 2002. Looking for French-English translations in comparable medical corpora. *Journal of the American Medical Informatics Association*, 8(suppl):150–154.

- K Church and P Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16:22–29.
- Béatrice Daille, Éric Gaussier, and Jean-Marie Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th COLING*, pages 515–521, Kyoto, Japan, August.
- Stéfan J. Darmoni, Éric Jarrousse, Pierre Zweigenbaum, Pierre Le Beux, Fiammetta Namer, Robert Baud, Michel Joubert, Huguette Vallée, Roger A. Côté, Antoine Buemi, Didier Bourigault, Gaele Recourcé, S. Jeanneau, and Jean-Marie Rodrigues. 2003. Extending the French part of the UMLS. In Mark Musen, editor, *Proceedings AMIA Annual Fall Symposium 2003*, page 824, Washington, DC, November. AMIA. (poster).
- P Fabry, R Baud, P Ruch, C Despont-Gros, and C Lovis. 2005. Methodology to ease the construction of a terminology of problems. *Int J Med Inform.*
- Éric Gaussier. 1998. Flow network models for word alignment and terminology extraction from bilingual corpora. In Christian Boitet, editor, *Proceedings of the 17th COLING*, Montréal, Canada, 10–14 August.
- L S Larkey and W B Croft. 1996. Combining classifiers in text categorization. In *Proceedings of SIGIR*, pages 289–297. ACM Press, New York.
- I. Dan Melamed. 2000. Bitext maps and alignments via pattern recognition. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*. Springer.
- Magnus Merkel, M Petterstedt, and Lars Ahrenberg. 2003. Interactive word alignment for corpus linguistics. In *Proceedings Corpus Linguistics*.
- Aurélié Névéol and Sylwia Ozdowska. 2005. Extraction bilingue de termes médicaux dans un corpus parallèle anglais/français. In *Proceedings EGC'05*.
- A Patry and Philippe Langlais. 2005. Paradocs: un système d'identification automatique de documents parallèles. In Michèle Jardino, editor, *Proceedings of TALN 2005 (Traitement automatique des langues naturelles)*, pages 223–232, Dourdan, June. ATALA, LIMSI.
- Philip Resnik and N.A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics*, 29:349–380. Special Issue on the Web as a Corpus.
- W Stolz. 1965. A probabilistic procedure for grouping words into phrases. *Language and Speech*, 8:219–235.
- D Widdows, B Dorrow, and C.K. Chan. 2002. Using parallel corpora to enrich multilingual lexical resources. In *Proceedings LREC*, pages 240–244, Las Palmas, Spain, May. ELRA.
- Dekai Wu. 2000. Bracketing and aligning words and constituents in parallel text using stochastic inversion transduction grammar. In Jean Véronis, editor, *Parallel Text Processing: Alignment and Use of Translation Corpora*. Springer.