Procesamiento del Lenguaje Natural, Revista nº 47 septiembre de 2011, pp 129-136 recibido 03-05-2011 aceptado 23-05-2011

Parallel corpus alignment at the document, sentence and vocabulary levels*

Alineación de corpus paralelo a nivel de documento, de oración y de vocabulario

Rogelio Nazar
IULA, Universitat Pompeu Fabra
Roc Boronat 138, 08018, Barcelona
rogelio.nazar@upf.edu

Resumen: Este artículo presenta un algoritmo independiente de lengua para la alineación de corpus paralelo a nivel de documento, de oración y de vocabulario, tomando como única fuente de información el mismo corpus a alinear. La entrada es un conjunto de documentos escritos en dos lenguas desconocidas A y B, donde cada documento en la lengua A tiene su correspondiente traducción a la lengua B. El problema consiste en: 1) dividir el conjunto de documentos en las dos lenguas; 2) alinear a nivel de documento: determinar qué documento en la lengua A es el original (o la traducción) de cada documento en la lengua B; 3) alinear a nivel de oración: determinar qué oración en el original corresponde a qué oración en la traducción y 4) alinear a nivel del vocabulario: determinar qué palabra en una lengua es equivalente a cada palabra en la traducción. El algoritmo es iterativo, ya que utiliza el vocabulario bilingüe resultante para realinear el corpus. La evaluación en inglés, castellano y francés muestra resultados competitivos en todos los niveles.
Palabras clave: Alineación de corpus paralelo, Extracción de información, traducción automática

Abstract: This paper presents a language independent algorithm for the alignment of parallel corpora at the document, sentence and vocabulary levels using the to-be aligned corpus itself as the only source of information. The input is a set of documents written in two unknown languages A and B, where every document in language A has its corresponding translation into language B. The problem thus consists of: 1) dividing the set of documents in the two languages; 2) aligning at the document level to determine which document in language A is the original (or translation) of each document in language B; 3) aligning at the sentence level to determine which sentence in the original corresponds to each sentence in the translation and 4) aligning at the vocabulary level to determine which word in one language is equivalent to each word in the translation. The algorithm is iterative, using the resulting bilingual vocabulary to re-align the corpus. Evaluation figures in English, Spanish and French show competitive results at all levels of the alignment.
Keywords: Parallel Corpus Alignment, Information Extraction, Machine Translation

1 Introduction

The days of parallel corpus processing probably started back in 1822, with the deciphering of the Rossetta stone by Prof. Jean-François Champollion (Véronis, 2000) and might have already started to end with the appearance of a series of works on extraction of bilingual vocabularies not from parallel but from monolingual corpora (Fung, 1995; Rapp, 1999). At present, parallel corpus alignment is a well established field (see Section 2) and it continues to be the best method to obtain a clean bilingual vocabulary from corpus. If various authors have started to explore new ways of acquiring bilingual vocabularies, it is not because of the limitations in the precision of the vocabulary alignment of parallel

* This research was funded by Project RICOTERM3 (Spanish Ministry of Science and Innovation, HUM2007-65966-C02-01/FILO) lead by Mercè Lorente. Many thanks also to the reviewers for their comments and to Aaron Feder for proofreading.

ISSN 1135-5948

corpus, but because parallel corpora of specialized domains are still not always easy to find, even today with the never ending availability of parallel corpora on the web (Resnik, 1999).

Despite the eventual difficulties in acquiring a parallel corpus of a specific domain, the fact is that parallel corpora of different domains are currently available. Think of the archives of international organizations such as the UN, the IMF or the EU. Having the possibility to align the parallel corpora produced by these organizations without taking into account the specific characteristics of each language or the coding they used would mean obtaining lexical resources of great value at zero cost. In addition, a language independent methodology would also be useful for the alignment of corpus of less resourced languages.

Overall, this paper presents a system that can be used to obtain a parallel corpus alignment at all levels starting from a set of documents written in two unknown languages, each document being the translation (or the original, an irrelevant difference for the purpose of this paper) of another document of the same set. The general process consists thus of the following operations, carried out sequentially: **1**) to separate the documents of the set in the two languages; **2**) to align the languages at the document level to determine which document is the translation of the other; **3**) to align them at the sentence level to determine, within each pair of original-translation documents, which sentence (or sentences) in the original corresponds to each sentence (or sentences) in the translation; **4**) to align them at the vocabulary level to identify the correspondences between the words of the two languages; **5**) To start again from step *2* introducing the bilingual vocabulary obtained in *4* as an additional parameter.

The paper is organized as follows: Section 2 is devoted to previous work on the subject of parallel corpus alignment. Section 3 presents a basic sketch of the present algorithm and in Section 4 this algorithm is evaluated with the alignment of parallel corpora in English, French and Spanish. Section 5 presents the conclusions and Section 6 draws some lines of future work.

2 State of the art

In general, the literature on statistical parallel corpus alignment has focused on the sentence and vocabulary levels. The number of papers in this field rapidly started to grow since the late eighties, after the seminal work of Kay and Röschesein (1988), who first introduced the idea of iterating the process of the alignment at the sentence level with the results of the alignment at the vocabulary level, thus mutually reinforcing the certainty of both alignments. Gale and Church (1991a) presented the co-occurrence statistics to obtain bilingual vocabularies and later (1991b), as Brown et al. (1991), they showed that there exists a correlation in length (in characters or words) in the equivalent sentences of a parallel corpus.

Many more coefficients were introduced later, such as the one based on the intuition that words in different languages which are orthographically similar (cognates) can be useful for the alignment (Church, 1993; McEnery and Oakes, 1995). A more geometric based intuition is that of Melamed (2000), who explains that if the alignment of a parallel corpus is represented in a plane where each position of the string characters in both texts is presented in the axes X and Y, then one can imagine the expected alignment function close to the the main diagonal, a straight line from the origin of slope close to $+1$.

Recently, different authors have focused on combining information from both sentence and vocabulary alignments to reinforce certainty and increase precision in subsequent passes of the aligner, as Moore (2002), who combined this information with variations of the series of IBM Models for statistical machine translation (Brown et al., 1993) and others who continued developing this idea, such as Varga et al.'s (2005) *Hunalign* and Braune & Fraser (2010). Another system based on IBM models is $Giza++$ (Och and Ney, 2000), for alignment at the word level.

Systems for sentence and vocabulary alignment have achieved high accuracy (over 98% in some cases) and therefore the race for precision is presumably over. At this juncture, further developments in the field should focus on novel approaches by integrating solutions to different problems, facing the challenge of computational efficiency or the economy of resources. While most authors working on parallel corpus alignment focus either

on sentence or vocabulary alignment, no research on integrating systems has been carried out. Some tools do integrate different levels of the analysis (Hiemstra, 1998; Simões and Almeida, 2003; Tiedemann, 2006), however the present proposal integrates solutions for all levels of the alignment with a new, simple and computationally efficient method, without explicit linguistic information and with a state-of-the-art level of precision.

3 Description of the algorithm

As already mentioned in the introduction, the input of this algorithm is a set of documents written in two unknown languages and the process consists of the following phases: separating the set of documents in the two languages (Section 3.1), aligning documents with their corresponding translations (Section 3.2.), aligning the sentences inside each pair of original-translation documents (Section 3.3.) and, finally, generating a bilingual vocabulary (Section 3.4.). The idea is that the process can start and finish without human intervention. However, in a real life application, a user may supervise the process and correct eventual errors at each step.

3.1 Separation of the set of documents in two languages

The problem of separating a set of documents written in two unknown languages is addressed taking advantage of the fact that documents written in the same language will have a number of vocabulary units in common no matter how different the topics of the two documents are. This set of overlapping units constitutes the most frequent part of the vocabulary. As a consequence, it is possible to cluster documents according to the number of frequent vocabulary units they have in common. This clustering can be formalized in two simple steps: 1) Select the largest document in the collection, name it D_a, and sort the vocabulary of this document by decreasing frequency order and 2) Place in a set A, together with D_a, all documents of the corpus which have at least three vocabulary units in common within their list of ten most frequent units and the ten most frequent units in D_a.

If the two languages are very similar (as is the case with French and Spanish) but there is a strong assumption about the fact that the parallel corpus is almost-perfect (meaning that we can rest assure that half of the documents are in one language and the other half in the other) then one can set the similarity threshold on half of the documents, or at some point between the half and such threshold.

pair	l	lNm	sim	voc	bvoc	mean
A_i, B_j	...	...	...	...	...	...

Table 1: Matrix for the comparison of pairs of documents in two languages

3.2 Alignment at the document level

With the result of the previous process at hand, in this step the algorithm builds a matrix of the correspondences between each original document and its translation. This matrix (shown in Table 1) includes a series of coefficients (*l*, *lNm*, *sim*, *voc*, *bvoc*, all explained in this section) which are calculated during the pairwise comparison of the documents in both languages. Thus, for each document in language A, we obtain a list of translation-candidate documents in language B, ordered by a score which is the mean of the values obtained for each individual coefficient.

Let us now define the coefficients listed in Table 1.

Coefficient *l*: This coefficient compares the length of the documents measured in number of characters, as Gale and Church (1991b) did to align at the sentence level. In this case, we assume that the original document and the translation have a similar size. As expressed in Equation 1, the function *length* refers to the length of a document in characters. In order to make up for the difference in natural redundancy in both languages (for instance, it takes more space to say the same in Spanish than in English), if a pair of documents obtain a value greater than or equal to .75, then this value is automatically adjusted to 1.

$$l(i,j) = \frac{\min(length(i), length(j))}{\max(length(i), length(j))} \qquad (1)$$

Coefficient *lNm*: This coefficient is based on the same idea as the previous one, however instead of comparing the length of the documents themselves, it compares the length of the titles of the documents (or, more

precisely, the names of the files), again assuming that if the titles of the documents are also translated, then it is to be expected that they will both have approximately the same length. It is also defined as Equation 1, only that in this case the letters i and j will denote file names instead of the text of the documents. Of course, this coefficient will not be useful when the names of the documents are arbitrary codes, as is the case in the experiments on Section 4.

Coefficient *sim*: Similarly to Coefficient *lNm*, this coefficient analyzes the names of the documents to measure orthographic similarity. The comparison is applied by transforming the names of the files into binary vectors which have sequences of two letters as dimensions. Similarity is measured using the Dice coefficient, expressed in Equation 2.

$$sim(I, J) = \frac{2|I \cap J|}{|I| + |J|} \quad (2)$$

Coefficient *voc*: As in the case of the first coefficient, this one is focused on properties of the documents themselves and not their names. It is based on the intuition that a genuine original-translation document pair has a high probability of having common vocabulary forms despite being written in different languages, and this probability is increased in the case of scientific or technical literature, due to the prevalence in those text genres of different symbols, numbers, acronyms and proper names. As expressed in Equation 3, this coefficient normalizes the number of vocabulary units in common by the number of different vocabulary units found in the larger of both documents.

$$voc(I, J) = \frac{|I \cap J|}{\max(|I|, |J|)} \quad (3)$$

Coefficient *bvoc*: This coefficient is somewhat similar to the previous one, but introduces a bilingual vocabulary. Recall from the introduction that a bilingual vocabulary is the result of the first iteration alignment. Thus the idea is that this coefficient is only applied after the second iteration when a bilingual vocabulary is already available or, alternatively, from the beginning in case a user is able to provide an external bilingual vocabulary. This coefficient is defined similarly as Coefficient *voc* (Equation 3), the only difference being that in this case the intersection between documents I and J is not the shared vocabulary as the same word types but as equivalent word types according to the bilingual vocabulary. Naturally, the larger the number of equivalent words in both documents, the more likely it is that one is the translation of the other.

3.3 Alignment at the sentence level

The general situation in the alignment at the sentence level shows many similarities with the previous alignment at the document level. However, in the sentence alignment there are new sources of information available, such as the position of the sentences in the document.

Coefficient *pos*: It is to be expected that the order of the sentences in one document is relatively the same as the one in the corresponding translation. Therefore, the first sentence of the source document will correspond to the first sentence in the translation and, similarly, the last sentence in the source will correspond to the last sentence in the translation. Under this assumption, there will be a correlation between the position of each sentence in the source and the position of the equivalent sentence in the translation. This expected correlation allows one to define a positional coefficient as indicated in Equation 4, where the symbols $P_{a,i}$ and $P_{b,j}$ represent the relative position of each sentence i in the original document a and the relative position of each sentence j in the translation.

$$pos(i, j) = \frac{\min(P_{a,i}, P_{b,j})}{\max(P_{a,i}, P_{b,j})} \quad (4)$$

The **Coefficient *sim***, which was already defined in Section 3.2., is also included in the matrix of sentence alignment, in this case not to be applied to names of files but to the sentences themselves. The purpose is to find cognates in a pair of candidate sentence pairs. The more cognates a pair has, the more likely it is that it will contain equivalent sentences. There is however a slight difference in the way this coefficient is implemented here, because in the case of the sentence comparison one cannot compare the whole sentences. Instead, it is necessary to separate both original and translated sentences into word vectors. This means that a comparison between sentences is equal to the pairwise comparison of the words of both sentences. Consequently, if two compared words surpass a Dice similarity value of .75, then both are considered

pair	l	pos	sim	num	voc	bvoc	mean
A_i,B_j	...	...	...	...	...	...	...

Table 2: Matrix for the alignment at the sentence level

pair	l	sim	coo	sum
A_i, B_j	...	...	...	...

Table 3: Matrix for the alignment at the vocabulary level

similar. The output of this coefficient is not just the output of the Dice coefficient but the number of similar words over the number of word types of the longer sentence.

In addition, a new **Coefficient *num*** is introduced with the purpose of detecting the presence of the same numbers in the candidate sentence pairs, assuming that numbers are language independent. The mechanics in which this coefficient is applied is exactly the same as the **Coefficient *voc***, and thus the same Equation 3 applies to define it.

Once all coefficients of this phase of the analysis have been introduced and defined, Table 2 presents the new matrix for the alignment at the sentence level including columns for the new positional (*pos*) and numeric (*num*) coefficients, with the rest of the coefficients already defined in Section 3.2.: the length ratio coefficient (l), the vocabulary overlap coefficient (*voc*) and, when available, the bilingual vocabulary overlap coefficient (*bvoc*). The best pair of aligned sentences a and b will be the one that obtains the highest mean of these coefficients, unless b has already been aligned before with other sentences and with a higher score. One cannot expect a perfect one-to-one sentence alignment because it is not a bijective function. As already mentioned, a sentence in the original can be translated as more than one sentence and vice-versa and, as a consequence, one needs a flexible system able to align a pair of sentences a and b even when b has already been aligned with a previous sentence of the original. As a solution, the algorithm tolerates multiple alignments for a sentence b up to an arbitrary threshold m (in these experiments, $m = 3$). In case sentence b has been previously proposed as the best alignment for other sentences more than m times, then it will be ignored in favor of the next candidate in the ranking.

3.4 Alignment at the vocabulary level

Once the corpus has been separated in languages and aligned at the document and sentence levels, it is now possible to elaborate a first version of the bilingual vocabulary which, in turn, will be used as an additional parameter in further iterations of the algorithm. The result of this vocabulary alignment is at the moment limited at the orthographic word, leaving the alignment at the expression level for future work (Section 6). As was the case in the alignment at the document and sentence levels, the alignment at the lexicon level implies a matrix for the comparison of words in different languages (Table 3). Two of the coefficients of this new matrix are common to the previous two matrices (Tables 1 and 2). The new Coefficient *coo*, defined below, adds information about the co-occurrence of the words in the aligned sentences, excluding hapax legomena and dis legomena. Another minor difference in this matrix is that the best candidate is not selected by the mean of the values obtained for each individual coefficient but for the sum of them. The motivation for this procedure is that the mean more often imposes an excessive penalization on correct candidates that obtained poor scores with l or *sim*.

Coefficient *coo*: This coefficient measures the statistical significance of the co-occurrence of two equivalent-candidate vocabulary units in two aligned sentences. Most studies in the extraction of bilingual vocabulary from parallel corpora relied in one way or another in statistics of co-occurrence of this type. The one applied in this paper (Equation 5) weights the co-occurrence of the candidate pair (i, j) over the total frequency of both units in the corpus.

$$coo(i,j) = \frac{f(i,j)}{\sqrt{f(i)} \cdot \sqrt{f(j)}} \tag{5}$$

As a result of the vocabulary alignment, for each vocabulary unit in the source language there is a list of equivalent candidates ordered by the score obtained in Table 3. The pairs of equivalent units are assembled using only the first candidate in the rank. In turn, each of the equivalent pair of units is assigned the same score, which means that pairs can also be ranked in order to retain only those which receive the highest ranking and are,

consequently, more reliable. The similarity threshold to consider a pair of units as equivalents is arbitrary. In the case of these experiments, the same threshold was used for the three languages.

4 Results obtained with the JRC-AQUIS corpus

This section describes the results obtained with the application of the methodology described in Section 3 to a fragment the JRC-AQUIS corpus (Steinberger et al., 2006). This is a collection of parallel corpora consisting of legislative text of the European Union in the languages of most of the member states. For the purpose of this experiment, only a fragment of the corpus was used, consisting of text in English, French and Spanish produced in 1995. The sample comprises 345 documents (approximately 800,000 word tokens) per language. All XML code and sentence alignment provided by Steinberger et al. was eliminated, leaving only plain text files.

Subsections 4.1., 4.2., 4.3. and 4.4 describe, respectively, the results obtained during the phases of language recognition, document alignment, sentence alignment and vocabulary alignment. Given the iterative nature of this algorithm, different numbers of precision and recall are reported, according to the performance achieved with each iteration. The reported results are obtained after five iterations, however no significant improvement is observed after the third iteration. The average time needed for the whole process (with its five iterations) was around five hours per pair of languages, using a single node of an Altix ICE 8200 machine (16 Gb of RAM) running on Linux.

4.1 Results of the separation of documents by language

Table 4 shows the results of the separation of documents by language in the pairwise comparison of the selected languages. This part of the process is not iterated. The results are in most cases correct, except a few cases in the French-Spanish pair, which can be explained by the similarity between the two languages.

	English	French
French	100%	—-
Spanish	100%	97%

Table 4: Precision in the separation of documents by language

	English	French
French	98%	—
Spanish	98%	86%

Table 5: Precision in the alignment at the document level (first iteration)

4.2 Results of the alignment at the document level

The results of the alignment at the document level are shown for each of the pairs of languages in Table 5. Those are the figures obtained after the first iteration. The results of the rest of the iterations up to five are almost identical. In the case of the Spanish-French pair, results are of course slightly worse because of the errors made on the previous step.

4.3 Results of the alignment at the sentence level

The sentence alignment was evaluated with random samples from the set of pairs of documents that were correctly aligned at the document level. Ten pairs of documents of each pair of languages were manually evaluated. The percentage of precision of the sentence alignment expressed in Table 6 represent the total proportion of the sentences correctly aligned for each set of 10 document pairs. The sample of documents was not large enough to show any evidence of progressive amelioration of the results after each iteration. Such progress, however, is evidenced by the improvement that can be appreciated in the alignment of the vocabulary, shown in the following section.

4.4 Results of the alignment at the vocabulary level

The last phase of the analysis is the extraction of bilingual vocabulary, which, as al-

	English	French
French	98%	—
Spanish	98%	98%

Table 6: Precision in the alignment at the sentence level (first iteration)

ready stated, consists of a list of vocabulary equivalences at the orthographic word level and not at the multiword expression level (see Section 6). Given that the results of this phase are also affected by the availability of the bilingual vocabulary, results are reported after one to five iterations, including the number of entries in the resulting vocabulary and the precision measured as the proportion of correct alignments.

As can be observed in Figure 1, the major increment in the number of entries is produced between the first and the second iteration. The growth diminishes in the next iteration until being almost null after the fourth iteration. Manual examination of the extracted bilingual vocabulary showed figures of precision over 98% in all cases. Recall is more difficult to evaluate because one cannot know how many units of the vocabulary should be aligned. By definition, leaving aside hapax legomena and dis legomena excludes half of the vocabulary. From the remaining units, it is only possible to estimate recall under the (wrong) assumption that every unit in the source has a corresponding translation. Considering that there are approximately 6,700 word types of frequency >= 3 on average per language, then recall approximates to 60%.

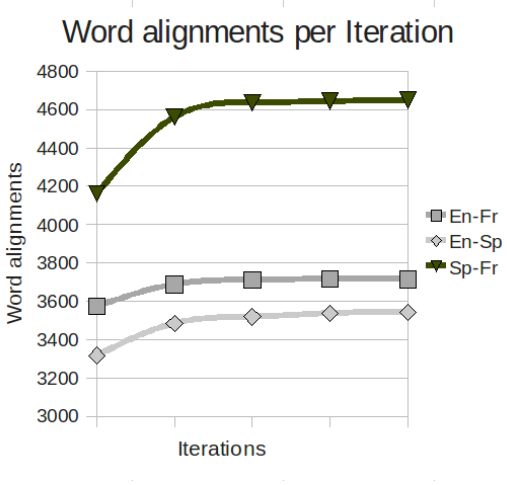

Figure 1: Results of the alignment at the vocabulary level

5 Conclusions

This article has presented a statistic algorithm for the alignment of parallel corpora at all levels. The empirical evidence obtained suggests that this algorithm can be applied to corpora of different domains and languages, with an output of sufficient quality to be used in real life scenarios, provided that there will be human post-editing. The time and effort that can be saved in the alignment of a raw corpus with this algorithm contrasts with its almost null cost of application, given that it needs no external resources.

Results of the separation of documents by language and the document alignment, as well as the sentence level, have been satisfactory. The alignment at the vocabulary level shows better precision than recall, however this is a desirable situation in this task, where achieving maximum recall is not a key issue. What is important in bilingual vocabulary extraction is that the information extracted is correct. If more units are needed, it is always possible to process more corpora. A problem that still has to be addressed is fertility, or the fact that a unit in one language can translate into a multiword expression, and this problem is reserved for future work (next Section).

6 Future work

This paper has placed considerable emphasis on the claim of language independence. However, as other authors warn (Choueka, Conley, and Dagan, 2000) these kinds of algorithms should be evaluated with totally unrelated pairs of languages such as Hebrew-English, Arab-English, or Chinese-English, in order to support that claim. Thus, replicating experiments in more languages should be the first line of future work. There are also ideas to improve the algorithm. For instance, it is likely that the performance of the sentence alignment would be better if it were done twice in both directions from one language to the other, however at the expense of duplicating the computational effort. With respect to the vocabulary alignment, future work will be devoted to alignment at the multiword expression level, which is essential for the parallel corpus alignment of specialized domains. There are many possibilities for this task, even when considering only statistic and language independent strategies. For instance, one can think of weighted n-grams as sequences of words with a significant frequency of occurrence. Other clues that can be used for multiword expression alignment are the total frequency of the units in the corpus and the document frequency, both of

which are expected to be similar in equivalent expressions.

References

Braune, F. and A. Fraser. 2010. Improved unsupervised sentence alignment for symmetrical and asymmetrical parallel corpora. In *Proc. of the the 23rd COLING International Conference*, pages 81–89, Beijing (China).

Brown, P., V. Della Pietra, S. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.*, 19:263–311, June.

Brown, P.F., J. C. Lai, and R. L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proc. of the 29th Annual Meeting of the ACL*, pages 169–176, Berkeley.

Choueka, Y., E. Conley, and I. Dagan. 2000. A comprehensive bilingual word alignment system. application to disparate languages: Hebrew and english. In *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer, pages 69–96.

Church, K.W. 1993. Charalign: a program for aligning parallel texts at the character level. In *Proc. of the 31st Annual Meeting of the ACL*, pages 1–8, Columbus, Ohio.

Fung, P. 1995. Compiling bilingual lexicon entries from a nonparallel english chinese corpus. In *Proc. of the Third Workshop on Very Large Corpora*, pages 173–183.

Gale, W. A. and K.W. Church. 1991a. Identifying word correspondences in parallel texts. In *Proc. of the DARPA Workshop on Speech and Natural Language*, pages 152–157.

Gale, W. A. and K.W. Church. 1991b. A program for aligning sentences in bilingual corpora. In *Proc. of the 29th Annual Meeting of the ACL*, pages 177–184, Berkeley.

Hiemstra, D. 1998. Multilingual domain modeling in twenty-one: automatic creation of a bi-directional translation lexicon from a parallel corpus. In *Proc. of the eightth CLIN meeting*, pages 41–58.

Kay, M. and M. Röschesein. 1988. Text-translation alignment. Technical report, Xerox Palo Alto Research Center.

McEnery, A. M. and M. P. Oakes. 1995. Sentence and word alignment in the crater project: methods and assessment. In *Proc. of the EACL-SIGDAT Workshop: from texts to tags, Issues in Multilingual Language Analysis(ACL)*, pages 77–86, Dublin, Ireland.

Melamed, D. 2000. Pattern recognition for mapping bitext correspondence. In *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer, pages 25–47.

Moore, R. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proc. of the 5th Conference of the Association for Machine Translation in the Americas on Machine Translation: From Research to Real Users*, AMTA '02, pages 135–144, London, UK, UK. Springer-Verlag.

Och, F. J. and H. Ney. 2000. Improved statistical alignment models. pages 440–447, Hongkong, China, October.

Rapp, R. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proc. of 37th Annual Meeting of the ACL*, pages 519–526.

Resnik, P. 1999. Mining the web for bilingual text. In *Proc. of 37th Annual Meeting of the ACL*, pages 527–532.

Simões, A. and J. Almeida. 2003. Natools - a statistical word aligner workbench. *Procesamiento del Lenguaje Natural*, 31:217–226, September.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proc. of the 5th LREC International Conference*.

Tiedemann, J. 2006. Isa ica - two web interfaces for interactive alignment of bitexts. In *Proc. of LREC 2006*, Genova (Italy).

Varga, D., L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. 2005. Parallel corpora for medium density languages. In *Proc. of the RANLP 2005*, pages 590–596.

Véronis, J. 2000. From the rosetta stone to the information society: A survey of parallel text processing. In *Parallel Text Processing: Alignment and Use of Translation Corpora*. Kluwer, pages 1–24.