Automatic induction of translation lexicons from aligned parallel corpus

Helena de M. Caseli[1], **Maria das Graças V. Nunes**[1]

[1]NILC – ICMC – University of São Paulo
CP 668P – 13.560-970 – São Carlos – SP – Brazil

{helename,gracan}@icmc.usp.br

Abstract. *Translation lexicons are one of the most important linguistic resources for machine translation. However, this bilingual set of word and multiword correspondences requires a lot of manual work to be built. This paper describes a method to automatically build translation lexicons by extracting knowledge from PoS-tagged and lexically aligned parallel corpora. Preliminary experiments were carried out on Brazilian Portuguese, Spanish and English parallel texts. The results showed that 85% of* pt–es *and 89% of* pt–en *entries are plausible correspondences. These results were obtained taking into consideration only the classes of entries which achieved the best results.*

1. Introduction

Two of the main challenges on natural language processing (NLP) are (1) the production, maintenance and extension of computational linguistic resources and (2) the integration of these resources into NLP applications.

In an attempt to overcome these challenges, several methods have been proposed to automatically build a variety of linguistic resources such as translation grammars [Carl 2001, Menezes and Richardson 2001] and translation lexicons [Wu and Xia 1994, Fung 1995, Gómez Guinovart and Sacau Fontenla 2004, Koehn and Knight 2002, Langlais et al. 2001, Schafer and Yarowsky 2002].

In line with [Gómez Guinovart and Sacau Fontenla 2004], this paper presents a method to automatically build a translation lexicon based on alignments produced by an automatic lexical aligner. This induced translation lexicon is more than just a list of source and target word equivalences. It is a set of bilingual word and multiword entries enriched by morphological and translation direction information. Such lexicon is an essential resource for transfer-based machine translation systems.

The experiments described in this paper are part of a larger project, ReTraTos. ReTraTos project aims to induce linguistic knowledge useful for machine translation – transfer rules and translation lexicons– for Brazilian Portuguese (pt) and its translations to other two languages: Spanish (es) and English (en). This paper focuses on translation lexicons specifically.

This paper is organized as follows. Section 2 presents related work on automatic building of translation lexicons. Section 3 summarizes the pre-processing tasks performed on the parallel corpora used for inducing the lexicons. The proposed induction method and the experiments carried out with pt–es and pt–en language pairs are described in Sections 4 and 5 respectively. This paper ends with some conclusions and proposals for future work (Section 6).

2. Related work

A translation lexicon is usually a side-product of a lexical alignment process. An automatic lexical aligner is a tool for finding correspondences between words, and sometimes multiword units, in parallel texts.

Several automatic lexical aligners have been proposed –among others, [Brown et al. 1993], [Och and Ney 2000] and [Caseli et al. 2005]– based on different alignment criteria such as statistics (e.g., co-occurrence frequency) and similarity (e.g., cognate measures).

In [Wu and Xia 1994], an English–Chinese lexicon was automatically induced by means of training a variant of the statistical model described in [Brown et al. 1993]. This model was trained on a large corpus (about 3 million words) resulting in a set of around 6,500 English words (on average 2.33 possible Chinese translations for each English word). Evaluation of a random set of 200 words showed an accuracy lying between 86.0% (complete automatic process) and 95.1% (manual correction).

By contrast, the method proposed by [Fung 1995] uses a non-aligned Chinese–English parallel corpus (with about 5,760 English words) to induce bilingual entries for nouns and proper nouns based on co-occurrence positions. Three judges evaluated the best induced entries (23.8%) and the average accuracy was 73.1%.

Other approaches have been also proposed in the literature. [Koehn and Knight 2002] proposes to build a translation lexicon from unrelated monolingual corpora. [Langlais et al. 2001] builds translation lexicons based on simple distributional properties of n-grams and little linguistic knowledge. [Schafer and Yarowsky 2002] combines two already existing translation lexicons to make a third one using one language as a bridge.

Following the approach taken by [Gómez Guinovart and Sacau Fontenla 2004], this paper proposes a method to induce translation lexicons which is based on alignments produced by an automatic lexical aligner. To our knowledge, no studies have yet been carried out to automatically build translation lexicons for Brazilian Portuguese.

3. Corpora pre-processing

The experiments described in this paper were carried out using two parallel corpora. One corpus consists of 18,236 pairs of `pt`–`es` parallel sentences (translation examples) with 1,049,462 tokens (503,596 in `pt` and 545,866 in `es`). Another corpus consists of 17,397 pairs of `pt`–`en` parallel sentences with 1,026,512 tokens (494,391 in `pt` and 532,121 in `en`). Both corpora contain articles from the online version of a Brazilian scientific magazine, *Pesquisa FAPESP*.[1]

These sets of translation examples were PoS-tagged using two tools available in `Apertium` [Armentano-Oller et al. 2006]: a morphological analyser and a PoS tagger. The morphological analyser provides one or more analysis (lemma, lexical category and morphological inflection information) for each surface form based on a monolingual morphological dictionary. The PoS tagger chooses the best possible analysis based on a first

[1] *Pesquisa FAPESP* is available at `http://revistapesquisa.fapesp.br`. It contains parallel texts written in Brazilian Portuguese (original), English (version) and Spanish (version).

order hidden Markov model (HMM).[2]

In ReTraTos project, the morphological dictionaries available at Apertium were enlarged. The pt and en dictionaries were enlarged with entries extracted from Unitex[3] [Paumier 2006] dictionaries. The es dictionary was enlarged with entries from the linguistic data used in the Spanish–Catalan machine translation system InterNOSTRUM[4] [Canals-Marote et al. 2001] provided by the Transducens machine translation group from the University of Alicante. The morphological dictionaries for pt and es available at Apertium es–pt linguistic data package (version 0.9) were enlarged to cover 1,136,536 and 337,861 surface forms respectively. The en morphological dictionary available at Apertium en–ca linguistic data package (version 0.8) was enlarged to cover 61,601 surface forms.[5]

After PoS-tagging, the translation examples were lexically aligned using two different tools: LIHLA [Caseli et al. 2005] and GIZA++ [Och and Ney 2000]. The translation examples were aligned in both directions (source–target and target–source) and the resulting alignments were merged by means of the union algorithm proposed by [Och and Ney 2003].

The pt–es examples were word-aligned using LIHLA with 94.25% precision and 94.97% recall. The pt–en examples were word-aligned using GIZA++ with 90.47% precision and 92.34% recall. These results were obtained by comparing the automatic alignments of a small set of sentences (about 500 sentences) with manually produced (reference) alignments as described in [Caseli 2007].

Figure 1 shows an extract of a pt–es translation example in which each surface form (the word as it appears in the text) is followed by tagger's output (its canonical form and PoS tags) and the alignment produced by the lexical aligner (the position of correspondent token on the other side).

Alignments of omission category are indicated by 0, such as the alignment of the second es token *que* which does not have any correspondence in the parallel sentence. Multiword unit alignments are concatenated (with "_") positions of correspondent tokens, as in the alignment between the pt word *esteja* (the 5th source token) and two es words: *se* (the 6th target token) and *encuentre* (the 7th target token). This $1:2$ alignment forms a target multiword unit. Multiword units can also be formed by the PoS tagger such as the es multiword unit *Pese_a*. The tagger is also responsible for marking unknown words with a "*" like **piquiá*.

4. Inducing the translation lexicon

The induction process presented in this paper comprises the following steps: (1) the compilation of two translation lexicons, one for each translation direction (one source–target and another target–source); (2) the merging of these two lexicons; (3) the generalization of bilingual entries; and (4) the treatment of syntactic differences related to entries in which

[2]The open-source machine translation system Apertium, the linguistic data and documentation is available at http://www.apertium.org.

[3]http://www-igm.univ-mlv.fr/~unitex/.

[4]http://www.internostrum.com/.

[5]Initially the pt, es and en morphological dictionaries covered 128,772, 116,804 and 48,759 surface forms respectively.

pt	<s snum=87>Embora/Embora<cnjadv>:1 o/o<det><def><m><sg>:3 *piquiá/piquiá:4 não/não<adv>:5 esteja/estar<vblex><prs><p3><sg>:6_7 sob/sob<pr>:8 risco/risco<n><m><sg>:9 de/de<pr>:10 ser/ser<vbser><inf>:11 extinto/extinto<adj><m><sg>:11 ,/,<cm>:12 a/o<det><def><f><sg>:13 exploração/exploração<n><f><sg>:14 descontrolada/descontrolado<adj><f><sg>:15 pode/poder<vbmod><pri><p3><sg>:16 levar/levar<vblex> <inf>:17 ao/a<pr>+o<det><def><m><sg>:18 desaparecimento/desaparecimento<n><m><sg>:19 dessa/de<pr>+esse<det><dem><f><sg>:0 ...
es	<s snum=87>Pese_a/Pese_a<pr>:1 que/que<cnjsub>:0 el/el<det><def><m><sg>:2 *piquiá/piquiá:3 no/no<adv>:4 se/se<prn><pro><ref><p3><mf><sp>:5 encuentra/encontrar<vblex><pri><p3><sg>:5 bajo/bajo<pr>:6 riesgo/riesgo<n><m><sg>:7 de/de<pr>:8 extinción/extinción<n><f><sg>:9_10 ,/,<cm>:11 la/el<det><def><f><sg>:12 explotación/explotación<n><f><sg>:13 desmesurada/desmesurado<adj><f><sg>:14 puede/poder<vbmod><pri><p3><sg>:15 ocasionar/ocasionar<vblex><inf>:16 su/suyo<det><pos><mf><sg>:17 desaparición/desaparición<n><f><sg>:18 en/en<pr>:21 ...

Figure 1. An extract of a pt–es translation example

the value of the gender or number attribute has to be determined based on information that goes beyond the scope of this entry.

In the first step, the method looks for all possible translations in the target (source) sentence for each source (target) word (its lemma, PoS tags and attributes), in each translation example. This search is performed based on the lexical alignments (see section 3). If more than one word is found in one or both sides, the PoS information of these words is joined by the character "+", forming a multiword unit. At the end of this step, the method stores all possible translations for each source (target) word or multiword unit and their occurrence frequency.

The translations found in the pt–es corpus for the pt word ao and its variations in terms of gender (*ao*, *à*) and number (*aos*, *às*) are used here to illustrate this point. *Ao* is a concatenation of a preposition (pr) and a determiner (det) and its PoS attributes have four possible sets of values: NC+<def><f><pl> (*às*), NC+<def><f><sg> (*à*), NC+<def><m><pl> (*aos*) and NC+<def><m><sg> (*ao*).[6] As shown in Table 1, each of these sets has several possible translations.

The ambiguity exemplified in Table 1 is solved in the next step, which merges the two translation lexicons (one for each translation direction) built in the first step. The lexicons are merged by: (1) choosing the translation with the highest occurrence frequency; (2) setting the valid translation direction (source–target or target–source), if necessary[7]; and (3) applying a frequency threshold to constrain the creation of multiword unit entries. An entry involving more than one word on one or both sides will be created

[6] Attribute values for type (def, definite), gender (f, feminine and m, masculine) or number (pl, plural and sg, singular) are indicated between "<" and ">" characters. The empty attribute sequence is represented by "NC".

[7] A bilingual entry is valid in both translation directions if the correspondence that it represents is the most frequent in both directions. When the correspondence is the most frequent in one direction only, this direction has to be indicated.

Table 1. Possible translations of pt word *ao* and its variations of gender and number found in pt–es translation examples

source word and attribute values	target translations and frequencies
a+o/pr+det	
NC+<def><f><pl>	a+el/pr+det = 156
	NC+<def><f><pl> = 140
	NC+<def><m><pl> = 9
	...
	el/det = 46
	<def><f><pl> = 36
	<def><m><pl> = 4
	...
	...
NC+<def><f><sg>	a+el/pr+det = 678
	NC+<def><f><sg> = 612
	NC+<def><f><pl> = 56
	...
	...
NC+<def><m><sg>	a+el/pr+det = 908
	NC+<def><m><sg> = 880
	NC+<def><f><sg> = 22
	...
	...
NC+<def><m><pl>	a+el/pr+det = 230
	NC+<def><m><pl> = 222
	NC+<def><m><sg> = 6
	...
	...

only if it occurs at least n times ($n = 50$ in the experiments presented in this paper). This constraint reduces the effect of wrong multiword unit alignments since, for this alignment category, the error rate is fairly high (11% in pt–es and 16% in pt–en parallel corpora).

In the second step, the bilingual entries in Table 1 are reduced to the entries shown in Table 2. For example, after merging the two translation lexicons, we found that, for the last set of source attribute values in Table 1 (NC+<def><m><pl>), a+el is the best target translation only when translating from source to target language, but not the other way round.

The third step tries to generalize the attribute values in bilingual entries with the same translation direction by merging the different values. For example, the first two entries in Table 2 can be joined together since they differ only in the value of number attribute (pl and sg). The resulting entry is shown in Table 3 with the merged value for number attribute (pl|sg) and the sum of both frequencies. During translation, the best value for number attribute is determined by the MT system.

Finally, the fourth step deals with entries whose values of the gender or number attributes can not be determined by the information in the entry. This happens when the same word is valid for both gender or number attribute values in one language but renders

Table 2. The best translations of pt word *ao* and its variations of gender and number found in pt–es translation examples considering both translation directions

source word and attribute values	best translations and frequencies	direction
a+o/pr+det	a+el/pr+det	
NC+<def><f><pl>	NC+<def><f><pl> = 140	both
NC+<def><f><sg>	NC+<def><f><sg> = 612	both
NC+<def><m><sg>	NC+<def><m><sg> = 880	both
NC+<def><m><pl>	NC+<def><m><pl> = 222	source–target

Table 3. The best translations of pt word *ao* and its variations of gender and number found in pt–es translation examples after generalization

source word and attribute values	best translations and frequencies	direction
a+o/pr+det	a+el/pr+det	
NC+<def><f><pl|sg>	NC+<def><f><pl|sg> = 752	both
NC+<def><m><sg>	NC+<def><m><sg> = 880	both
NC+<def><m><pl>	NC+<def><m><pl> = 222	source–target

two different translations in the other language, one for each attribute value. In this step, for each word, the system looks for an entry which has the general value for either gender (mf) or number (sp) on one side and, on the other side, there is the merged value for either gender (f|m) or number (pl|sg). If such entry is found, the system replaces it with three entries according to the translation directions: one for each attribute value and another replacing the merged value with the value of gender (GD) or number (ND) to be determined.

For example, the es noun (n) *análisis* can be translated as both the singular pt noun *análise* and the plural pt noun *análises*. As shown in Table 4, three entries are built to deal with this specific number problem.

Table 4. Example of dealing with grammatical differences of number for the pt word *análise* and its variations on number

source word and attribute values	best translations	direction
análise/n	análisis/n	
<f><sg>	<m><sp>	source–target
<f><pl>	<m><sp>	source–target
<f><ND>	<m><sp>	target–source

5. Experiments and results

Two translation lexicons were induced. The first has 23,450 pt–es entries and was induced from 18,236 pt–es translation examples. The second has 19,191 pt–en entries and was induced from 17,397 pt–en translation examples.

5.1. Evalutation of the pt–es translation lexicon

The automatically induced pt–es translation lexicon has 23,129 single word and 321 multiword unit entries. This lexicon was first evaluated by automatically comparing it

with the translation lexicon used by Apertium (10,360 word entries and 928 multiword unit entries from es–pt linguistic data version 0.9). All bilingual entries were compared and classified as identical, new or different. The different and new entries were manually analyzed. Different entries are those with the same source side found in Apertium's lexicon but different target side, attribute values or translation direction. New entries are those whose source side is not found in Apertium's lexicon.

The automatic comparison showed that 13% of single word entries are identical in both lexicons. This percentage rose to 15% for multiword units. Around 23% of single word and 13% of multiword unit entries differ in some aspect. However, this does not mean that they are incorrect. The most important contribution relies on the number of new entries: 63% of word entries and 72% of multiword unit entries do not occur in Apertium's lexicon.

The new and the different entries for words and multiword units (86.53% of all induced entries) were divided into more specific classes in order to evaluate each new or different information separately. Within these classes, 52.84% of the entries should not be added to a translation lexicon without a careful manual revision. These entries were discarded due to one of three reasons. First, 36.86% of the entries were not PoS-tagged. Second, 15.95% of the entries represented the same information on Apertium's lexicon but with more attribute values. Third, 0.03% of the entries were probably wrong since their PoS were not the same found in Apertium's lexicon.

The remaining entries (33.67% of all induced entries) belong to one of six classes: 22.53% new (N), 4.17% new translation direction (NTD), 5.91% different (D), 0.85% more general translation direction (MGTD), 0.09% different translation direction (DTD) and 0.12% more general attribute values (MGAV). The MGTD and MGAV classes include entries whose values of translation direction and attribute values, in this order, are more general than those found in Apertium's entries.

Two human specialists in pt and es evaluated a random set of entries which consisted of 10% of all entries from these six classes (474 entries = 459 for single words and 15 for multiword units). These entries were classified as either plausible or implausible. This analysis has shown that many entries were classified as implausible due to tagging errors. For example, the entry which represents the correspondence between the pt adjective *requintado* and the es verb (past participle) *sofisticado* was classified as implausible. This correspondence would be plausible if the tagger had tagged *sofisticado* as an adjective and not as a verb. In addition to PoS errors, wrong attribute values can also cause an entry to be classified as implausible.

Table 5 shows the results of manual analysis after the entries with tagging errors have been filtered. Due to the small number of multiword unit entries classified as MGTD, DTD or MGAV, no entry from these classes was selected to be manually evaluated. In fact, each judge evaluated 6% of entries with an overlap of 2% which was designed to measure the agreement between both judges by means of the kappa measure [Carletta 1996]. The value of kappa was 0.63 (good agreement). This figure reflects the judges' disagreement on 15.61% of single word entries and 6.06% of multiword unit entries.

These results point to one problematic class for single word entries (NTD) and two for multiword unit entries (N and D). A lot of word entries classified as new translation di-

Table 5. Percentage of plausible pt–es entries according to manual analysis

Classification	Words(%)	Multiwords(%)	All(%)
NEW	**76.12**	**52.63**	**75.20**
new (N)	79.30	42.86	78.07
new translation direction (NTD)	57.35	80.00	58.90
DIFFERENT	**77.39**	**0.00**	**74.65**
different (D)	72.48	0.00	69.30
more general translation direction (MGTD)	94.44	–	94.44
different translation direction (DTD)	100.00	–	100.00
more general attribute values (MGAV)	100.00	–	100.00
TOTAL	**76.40**	**41.67**	**75.08**

rection are plausible correspondences in the context of lexical alignment, but implausible in the context of a translation lexicon. This is the case of the entry that sets the correspondence between the es noun *organizadoras* and the pt noun *organização*. Most problems on multiword unit entries are due to incomplete correspondences such as that between the es multiword expression *como consecuencia* and the pt single word *conseqüência* (the plausible correspondence, in this case, would be *consecuencia* and *conseqüência*).

As regards the results of the manual analysis, we can conclude that 75% of pt–es entries are plausible: 76% of word entries and 42% of multiword unit entries. The low percentage of plausible multiword unit entries is due to the high error rate in the lexical alignment of $n : m$ instances (11.19%). If we consider only the entries from the classes which achieved the best results (word entries from N, D, MGTD, DTD and MGAV classes and multiword entries from NTD class) plus the identical ones, the resulting set of 9,930 entries is expected to reach 85% accuracy.

5.2. Evalutation of the pt–en translation lexicon

The automatically induced pt–en translation lexicon has 15,949 single word and 3,242 multiword unit entries. The automatic comparison of this induced translation lexicon with another manually built, as has been done for pt–es, was not possible since such manual pt–en lexicon was not available. The pt–en entries were first evaluated by automatically classifying them as equal, incomplete (not PoS-tagged) or different comparing source and target sides. The incomplete entries were discarded and the equal and different entries were manually analyzed.

From automatic classification, four classes (53.71% of all induced entries) were selected to be manually evaluated: 7.75% same source and target category and attribute values (SCAV), 27.77% more specific attribute values (MSAV), 0.90% more general attribute values (MGAV) and 17.29% different attribute values (DAV). The MSAV and MGAV classes include entries whose source attribute values are more specific and more general, in this order, than target ones.

Two human specialists in pt and en evaluated a random set of entries which consisted of 10% of all entries from these four classes (1,030 entries = 865 for single words and 165 for multiword units). These entries were classified as either plausible or implausible. Again, an overlap of 2% was designed to measure the agreement between both judges by means of the kappa measure. The value of kappa was 0.48. This figure reflects

Table 6. Percentage of plausible pt–en entries according to manual analysis

Classification	Words(%)	Multiwords(%)	All(%)
EQUAL	**94.74**	**53.06**	**80.56**
same category and attribute values (SCAV)	94.74	53.06	80.56
DIFFERENT	**81.08**	**30.97**	**74.60**
more specific attribute values (MSAV)	92.25	50.00	91.44
more general attribute values (MGAV)	90.00	60.00	80.00
different attribute values (DAV)	56.17	27.55	47.75
TOTAL	**82.59**	**37.65**	**75.44**

the judges' disagreement on 16.18% of single word entries and 45.45% of multiword unit entries. The high disagreement on multiword unit entries is due to one of the judges pointed out that most of these entries should be applied to just one translation direction, rather than to both as indicated by these entries.

Table 6 shows the results of manual analysis. As regards these results, we can conclude that 75% of pt–en entries are plausible: 83% of word entries and 38% of multiword unit entries. Again, the lower percentage of plausible multiword entries is due to high error rate in the lexical alignment of $n : m$ instances (15.71%). This is the case of the alignment between the pt word *jantar* and the en multiword expression *dinner already* (the plausible correspondence, in this case, would be *jantar* and *dinner*).

If we consider only the entries from the classes which achieved the best results (SCAV, MSAV and MGAV) the resulting set of 6,988 entries is expected to reach 89% accuracy. If we consider only the 6,338 single word entries from the classes which achieved the best results the accuracy increases to 93%.

6. Conclusions and future work

Preliminary experiments carried out on the pt-es and pt–en automatically induced translation lexicons brought out interesting results. These experiments were performed in two steps. First, we automatically compared (pt–es) or classified (pt–en) the induced entries. Second, two judges performed a manual analysis of random sets of entries.

The main contribution of the method presented in this paper is that it can be used by any MT system to automatically build their own translation lexicons. As future work, we aim to use the induced lexicons in a MT system together with other linguistic resources induced automatically by ReTraTos project.

Acknowledgements

We thank FAPESP, CAPES and CNPq for financial support. We also thank Mônica Martins, Élen Tomazela, Gema Ramírez-Sánchez, Carmen Dayrell and Transducens group in the University of Alicante for contributions to this work.

References

Armentano-Oller, C., Carrasco, R. C., Corbí-Bellot, A. M., Forcada, M. L., Ginestí-Rosell, M., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., and Scalco, M. A. (2006). Open-source Portuguese-Spanish machine translation. In *Proceedings of the VII PROPOR*, pages 50–59, Itatiaia-RJ, Brazil.

Brown, P., Della Pietra, V., Della Pietra, S., and Mercer, R. (1993). The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–312.

Canals-Marote, R., Esteve-Guillén, A., Garrido-Alenda, A., Guardiola-Savall, M., Iturraspe-Bellver, A., Montserrat-Buendia, S., Ortiz-Rojas, S., Pastor-Pina, H., Pérez-Antón, P., and Forcada, M. (2001). The Spanish-Catalan machine translation system interNOSTRUM. In *Proceedings of MT Summit VIII*, pages 73–76.

Carl, M. (2001). Inducing probabilistic invertible translation grammars from aligned texts. In *Proceedings of CoNLL-2001*, pages 145–151, Toulouse, France.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2):249–254.

Caseli, H. M. (2007). *Indução de léxicos bilíngües e regras para a tradução automática*. PhD thesis, ICMC–USP–São Carlos.

Caseli, H. M., Nunes, M. G. V., and Forcada, M. L. (2005). Evaluating the LIHLA lexical aligner on Spanish, Brazilian Portuguese and Basque parallel texts. *Procesamiento del Lenguaje Natural*, 35:237–244.

Fung, P. (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of ACL-1995*, pages 236–243.

Gómez Guinovart, X. and Sacau Fontenla, E. (2004). Métodos de optimización de la extracción de léxico bilingüe a partir de corpus paralelos. *Procesamiento del Lenguaje Natural*, 33:133–140.

Koehn, P. and Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In Association for Computational Linguistics, editor, *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pages 9–16, Philadelphia.

Langlais, P., Foster, G., and Lapalme, G. (2001). Integrating bilingual lexicons in a probabilistic translation assistant. In *Proceedings of MT Summit VIII*, pages 197–202, Santiago de Compostela, Spain.

Menezes, A. and Richardson, S. D. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In *Proceedings of the Workshop on Data-driven Machine Translation at 39th ACL*, pages 39–46.

Och, F. J. and Ney, H. (2000). Improved statistical alignment models. In *Proceedings of the 38th ACL*, pages 440–447, Hong Kong, China.

Och, F. J. and Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Paumier, S. (2006). *Unitex 1.2 user manual*. Université de Marne-la-Vallée.

Schafer, C. and Yarowsky, D. (2002). Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the 6th CoNLL, co-located with COLING-2002*, Taipei, Taiwan.

Wu, D. and Xia, X. (1994). Learning an English-Chinese lexicon from parallel corpus. In *Proceedings of the 1st AMTA*, pages 206–213, Columbia, MD.