

Improving Statistical Word Alignments with Morpho-syntactic Transformations

Adrià de Gispert¹, Deepa Gupta², Maja Popović³, Patrik Lambert¹,
Jose B. Mariño¹, Marcello Federico², Hermann Ney³, Rafael Banchs¹

¹ TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain

² ITC-irst, Centro per la Ricerca Scientifica e Tecnologica, Trento, Italy

³ Lehrstuhl für Informatik 6, RWTH Aachen University, Aachen, Germany

Abstract. This paper presents a wide range of statistical word alignment experiments incorporating morphosyntactic information. By means of parallel corpus transformations according to information of POS-tagging, lemmatization or stemming, we explore which linguistic information helps improve alignment error rates. For this, evaluation against a human word alignment reference is performed, aiming at an improved machine translation training scheme which eventually leads to improved SMT performance. Experiments are carried out in a Spanish–English European Parliament Proceedings parallel corpus, both in a large and a small data track. As expected, improvements due to introducing morphosyntactic information are bigger in case of data scarcity, but significant improvement is also achieved in a large data task, meaning that certain linguistic knowledge is relevant even in situations of large data availability.

1 Introduction

Word aligned corpora are useful in a variety of fields. An obvious one is automatic extraction of bilingual lexica and terminology [1]. Word sense disambiguation is another application [2], since ambiguities are distributed differently in different languages. Word aligned corpora can also help for transferring language tools to new languages. In Yarowsky and Wicentowski [3], text analysis tools such as morphologic analyzers or part-of-speech taggers are projected to languages where such resources do not exist. Kuhn [4] presents a study of ways for exploiting statistical word alignment for grammar induction.

In statistical machine translation (SMT), word alignment is a crucial part of the training process. In approaches based on words [5], phrases [6] or n-grams [7], the basic translation units are indeed extracted from statistical word alignment [8]. Some syntax-based SMT systems [9] also rely on word alignment to estimate tree-to-string or tree-to-tree alignment models.

Och and Ney [10] have shown that translation quality depends on word alignment quality

In this paper we study ways of improving alignment quality through the incorporation of morpho-syntactic information. This type of information has already

been used to enhance word alignment systems: Toutanova et al. [11] augmented a HMM statistical alignment model with POS tags data; Tiedemann [12] and de Gispert [13] computed system features based on POS tags, chunk labels or lemmas. Popović and Ney [14] used hierarchical lexicon structure enriched with German base forms and POS tags for the EM training of German-English alignments.

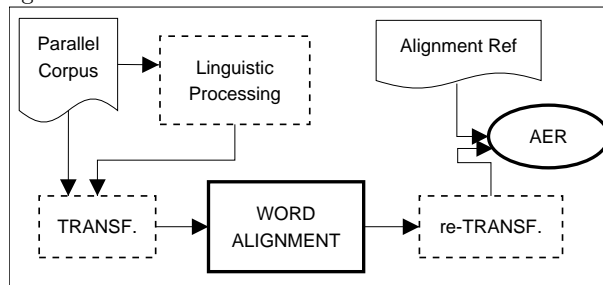
In the experiments described here, the alignment models remain purely statistic, whereas the training corpus is transformed so as to make the statistical alignment models task easier. Results are evaluated measuring the Alignment Error Rate against a manual reference (see section 3.2).

The organization of the paper is as follows. Section 2 presents the morphosyntactic data transformations that have been considered to improve alignment, whose results are shown and discussed in section 3. Finally, section 4 concludes and gives ideas of future work.

2 Morphosyntactic Corpus Transformations

With the goal of finding out which linguistic features are relevant for improving statistical word alignment, we have followed a corpus transformation approach, ie. data has been modified using morphosyntactic information before word alignment, as shown in the flow diagram in Figure 1.

Fig. 1. Experimental configuration to evaluate impact of using morphological information on word alignment.



Then, the obtained alignment of the transformed parallel corpus is mapped to the original sentence pairs in order to evaluate Alignment Error Rate against a manual reference. The same word alignment algorithm and configuration is used in all cases, therefore acting as a black-box.

In many cases, the corpus transformation can be seen as a classification from words to linguistically-enriched tokens, be it of all words or just some groups of words. However, we have also considered linguistically-motivated word order modifications, as well as combinations of both.

As most transformations are done on a word basis, aligned tokens can be directly substituted by original text after word alignment. In case some words are grouped in a single token before aligning, all internal links are introduced when writing back original text.

Now each of the transformations carried out leading to an independent experiment, is motivated and fully described.

2.1 Word Classifications

In general, word classifications aim at reducing data sparseness, by mapping some words to a unique token according to a certain criterion. In our case, criteria are based on the linguistic information provided by state-of-the-art language tools, in the particular case of processing the Spanish and English languages.

Base forms Also known as lemmas, base forms lack details on morphological derivation of the word (gender, number, tense, and so on) and only provide information on the head of the word. Therefore, they represent a meaning-bearing reduced version of each word, especially in the case of high morphological derivation, such as verbs, nouns or adjectives in Spanish. In English, verbs and nouns are also reduced by taking the base form, even though in lesser degree.

Stems Same as lemmatization, stemming is another method of word transformation which truncates inflected word forms by a single stem without morphological suffixes or derivations. However, a stemmer may not necessarily produce any meaning-bearing word form, whereas a lemmatizer returns the base form, usually associated with a dictionary citation of the given word form. Table 8 gives an example of stemming and lemmatization results illustrating the differences between the two processes.

Spanish Adjective Base Forms Spanish adjectives, in contrast to English, have gender and number inflections so that one base form can have four different full forms. For instance, the adjective “*bonito*” (beautiful/pretty) has four inflected forms (“*bonita*”, “*bonitas*”, “*bonito*”, “*bonitos*”). Therefore, reducing the inflection from the Spanish adjectives might simplify the process of word alignment between two languages. All Spanish adjectives are replaced with its base forms whereas the English corpus remains the same.

Reduced Spanish Verbs Spanish language has an especially rich inflectional morphology for verbs. Person and tense are expressed via suffix so that many different full forms of one verb exist, many of them without the corresponding equivalent in English. Therefore, reducing the POS information of Spanish verbs could be helpful for improving word alignments. Each verb has been reduced into its base form and reduced POS tag: parts of POS tag describing tense and/or mode which does not exist in English are removed. For example, the tag for the

subjunctive mode has been removed, and the two tags representing two types of the past tense are replaced with the unique past tense tag.

Lemma plus reduced Spanish POS Morpho-attributes As already mentioned, Spanish is morphologically richer than English. However, all inflected forms of Spanish are not relevant for translation into English. For instance, whereas Spanish adjectives may have four inflected forms, English adjectives have only one form. Therefore, it might be possible that all inflected forms of Spanish adjectives are not required for translation. Similar cases are possible to a limited extent with other words also, such as nouns, verbs, etc.

To handle this morphology-related problem of Spanish with respect to English, we can count for each Spanish part of speech (POS) tag which additional morphological attributes (morpho-attributes) do not affect the translation from Spanish to English. For this purpose, we extract bilingual lexicons from original word-based statistical word alignment for large training data from both directions (Spanish to English and English to Spanish), where each Spanish original word is replaced with its lemma plus morpho-syntactic tag. On this bilingual lexicons, entropy was calculated with respect to each morpho-attribute corresponding to each Spanish POS tag. As a result, Table 1 reports that irrelevant and relevant morpho-attributes corresponding to some Spanish POSs. Other Spanish POS (adverbs, conjunctions and interjections) have not been reported in the table as they do not convey enough morphological information. In case of some morpho-attributes for Spanish POS, the value of the entropy was not significantly reduced with respect to the value of the entropy considering only with lemma form. In this situation, we tried different combination of morpho-attributes for that POS. For instance, Table 1 reports relevant morpho-attributes for determiner are gender and number. We observed that for small data track, these morpho-attributes do not make significant effect on the translation. Therefore, in case of small data track, we have not provided this information with lemma form.

In general, Spanish words are replaced with lemma and its relevant POS tag information. The remaining ones are transformed into lemma forms in small as well as in large data (see Table 8 for example).

Table 1. Irrelevant & Relevant POS Morphological Attributes for Spanish.

POS	Irrelevant POS morpho-attributes	Relevant POS morpho-attributes
Verb	type (principal, auxiliary)	mode, time, person, number, gender
Noun	type (common, proper), gender, grade	number (singular, plural, invariable)
Adjective	type, grade, gender, number, function	–
Pronoun	person, possessor, politeness	type, gender, number, case
Determiner	type (demonstrative, possessive, etc.) person, possessor	gender, number
Preposition	type, form, gender, number	–

Full Verb Forms Undoubtedly, given a verb meaning, tense and person, each language *implements* each verbal form independently from the other language. For example, whereas the personal pronoun is compulsory in English unless the subject is present, this does not occur in Spanish, where the morphology of the verb expresses the same aspect.

Therefore, aiming at simplifying the work for the word alignment, another word classification strategy can be devised to address the rich variety of verbal forms. For this, we group all words that build up a whole verbal form (including pronouns, auxiliary verbs and head verb) into the lemma of the head verb. This is a knowledge-based detection taken using deterministic automata implementing a few simple rules. These rules require information on word forms, POS-tags and lemmas in order to map the resulting expression to the lemma of the head verb, as done in [13]. Examples of such mappings can be found in Table 2.

Table 2. Full verb forms are mapped to the lemma of the head.

English		Spanish	
full form → lemma		full form → lemma	
has been found	find	introdujeran	introducir
we will find	find	han cometido	cometer
do you think	think	dijo	decir
offered	offer	está haciendo	hacer
I am doing	do	haremos	hacer

2.2 Word Order Modification

It is commonly known that non-monotonicity poses difficulties for word alignment, not to mention for statistical machine translation. The more differences in word order between two languages, the more difficult to extract a good alignment and the more challenging the translation task is. Although English and Spanish exhibit a quite remarkable monotonicity (compared to other pairs such as English and Chinese), here we study two techniques, exploring the possible gain in alignment quality of reordering one language to make word alignment more monotone.

POS-based Reordering of Spanish Nouns and Adjectives Adjectives in Spanish are usually placed after the corresponding noun, whereas in English it is the other way round. Therefore local reordering of nouns and adjective groups might be helpful for monotonising word alignments between two languages. POS-based local reordering [15] has been used: each Spanish noun has been moved behind the correspondent adjective group. If there are two adjectives connected with a coordinate conjunction “and” or “or”, the noun is moved behind the whole group of words.

Noun–Adjective swapped realignment An alternative strategy consists of deciding which Spanish 'Noun + Adjective' structures need to be swapped according to classes extracted from an initial statistical word alignment in the original order, as introduced in [16].

Given this baseline alignment, we build up classes of nouns preceding the same adjectives and having crossed links⁴. The same classes can be extracted for the adjectives following the same nouns. From these classes, we filter out those pairs occurring less than 6 times or having a low crossed-link probability, ie. being more often monotonically linked.

Finally, we swap all remaining 'Noun + Adjective' belonging to seen pairs of classes, and realign, as we expect the increase in monotonicity to reduce the word alignment complexity and improve quality.

2.3 Combinations

Two types of combinations can be performed. On the one hand, one can combine two (or more) presented approaches to produce a new transformation. For example, any word order modification can be done together with stemming, base form substitution or full verb classification. Verb classification can also be combined with other transformation for all words outside the verb groups.

On the other hand, a new word alignment can be obtained from the combination via consensus of the different alignments generated by various transformations. Both these strategies have been tested in order to achieve the best alignment quality.

3 Experimental Framework

3.1 Corpus Description

Experiments have been carried out using the Spanish-English EPPS parallel corpus, which contains the debates proceedings of the European Parliament from 1996 to May 2005. In order to extract the linguistic information needed to perform the presented corpus modifications, we preprocessed the data as follows:

- English POS-tagging using freely-available *TnT* tagger [17].
- English lemmatization using *wmmorph*, included in the WordNet package [18].
- Spanish POS-tagging and lemmatization using *FreeLing* analysis tool [19].
- English and Spanish stemming using the Snowball stemmer⁵, which is based on Porter's algorithm.

Table 3 shows the main statistics of the parallel corpus used, including number of sentences, number of words, vocabulary and average sentence length for each language. The lower part of the table shows the statistics for the 1% division used in the small data track.

⁴ By crossed links, we mean that Spanish word in position n is linked to English word in position $m + 1$, and Spanish word in $n + 1$ is linked to English word in m .

⁵ <http://www.snowball.tartarus.org/>

Table 3. Parallel corpus statistics for large and small data tracks.

	sent	words	vocab.	avg len
English	1.28 M	34.9 M	106 k	27.2
Spanish		36.6 M	153 k	28.5
English 1%	13.4 k	366 k	16.3 k	27.4
Spanish 1%		385 k	22.4 k	28.8

3.2 Evaluation measures and manual reference

For evaluation, an ample set of bilingual sentences was aligned manually (see table 4), following a carefully defined procedure [20] by agreement of three manual reference alignments. 66.7% of reference alignment links are Sure whereas 33.3% are Possible. This alignment test set is a subset of the training data, both in the large and the small data tracks.

Table 4. Alignment test data statistics.

	sent	words	vocab.	avg len
English	400	11.7 k	2.7 k	29.1
Spanish		12.3 k	3.1 k	30.4

The alignment test data contain unambiguous links (called S or Sure) and ambiguous links (called P or Possible). If there is a P link between two words in the reference, a computed link (*i.e.* to be evaluated) between these words is acceptable, but not compulsory. On the contrary, if there would be an S link between these words in the reference, a computed link would be compulsory. In this paper, precision refers to the proportion of computed links that are present in the reference. Recall refers to the proportion of reference Sure links that were computed. The alignment error rate (AER) is given by the following formula:

$$AER = 1 - \frac{|\mathcal{A} \cap \mathcal{G}_S| + |\mathcal{A} \cap \mathcal{G}|}{|\mathcal{A}| + |\mathcal{G}_S|} \quad (1)$$

where \mathcal{A} is the set of computed links, \mathcal{G}_S is the set of Sure reference links and \mathcal{G} is the entire set of reference links.

3.3 Baseline Statistical Word Alignment

As word alignment core algorithm, GIZA++ [21] was used. Regarding model iterations, we use the $1^4H^34^4$ configuration (meaning 4 iterations of IBM model 1, 5 iterations of HMM model and 4 iterations of IBM model 4), which provides

the best AER for our task. During word alignment, we use 50 classes per language as estimated by 'mkcls', a freely-available tool along with GIZA++⁶.

Moreover, we always work with lowercase text before aligning, as this leads to a significant AER reduction when compared with the true-case text. Note that this configuration applies for all experiments that have been done.

3.4 Alignment results

Table 5. Word Alignment results for small-data task.

	Eng→Spa			Spa→Eng			Union		
	R_S	P_P	AER	R_S	P_P	AER	R_S	P_P	AER
baseline	63.10	77.11	30.34	64.12	80.21	28.38	73.37	69.43	28.77
base forms	66.37	83.50	25.75	68.06	83.72	24.69	73.93	75.01	25.51
stems	67.02	84.30	25.01	68.61	83.80	24.32	74.66	75.65	24.82
Spa Adj base	63.96	78.29	29.33	64.17	80.31	28.31	73.59	70.19	28.25
Spa V reduced	64.25	78.39	29.13	64.09	80.16	28.44	73.17	70.05	28.51
Spa lem+redPOS	64.36	80.63	28.06	64.51	79.08	28.70	73.71	70.76	27.87
full verbs	66.50	79.72	27.13	65.44	81.30	27.10	73.96	71.36	27.45
Spa N-A reord	63.44	77.27	30.08	64.57	80.39	28.04	73.40	69.68	28.61
N-A swap realign	63.63	77.41	29.91	64.27	80.00	28.38	73.43	69.59	28.65
verbs + stems	69.58	83.17	23.89	67.33	83.96	24.85	75.47	75.17	24.69

Table 6. Word Alignment results for large-data task.

	Eng→Spa			Spa→Eng			Union		
	R_S	P_P	AER	R_S	P_P	AER	R_S	P_P	AER
baseline	73.20	90.78	18.65	72.18	92.17	18.64	78.42	86.43	17.56
base forms	72.80	91.70	18.54	71.84	93.17	18.50	76.73	87.90	17.82
stems	73.56	92.40	17.79	72.72	93.78	17.68	77.81	88.94	16.74
Spa Adj base	73.01	90.78	18.77	72.40	92.47	18.39	78.30	86.70	17.50
Spa V reduced	73.07	90.69	18.77	72.07	92.22	18.70	77.97	86.43	17.80
Spa lem+redPOS	72.72	90.46	19.06	71.94	92.06	18.82	77.87	86.16	17.97
full verbs	74.27	90.77	17.85	73.03	93.31	17.56	78.60	87.37	16.97
Spa N-A reord	72.69	90.06	19.25	72.23	91.85	18.73	78.10	85.93	17.97
N-A swap realign	72.52	90.41	19.22	72.13	91.80	18.81	77.91	86.10	17.99
verbs + stems	74.74	91.83	17.14	73.23	93.84	17.23	78.36	88.82	16.42

Results with the 1% data set are shown in Table 5, where both directions and the symmetrization through union are evaluated. Each row refers to each of the corpus transformations presented.

⁶ See <http://www.fjoch.com> for details on both tools.

As it can be seen, both **base forms** and **stems** produce a very significant quality improvement, especially reflected in a more than 5 point absolute precision improvement in union alignment, whereas recall is also very high in these two cases for all alignment directions. It looks like their classifications reduce sparseness and help the word alignment algorithm perform better. This improvement is best in the case of stems.

Whereas '**Spa lem+redPOS**' transformation also achieves significant improvements in recall and precision for all directions, leading to an approximate 1 point AER reduction, improvements due to '**Spa Adj base**' and '**Spa V reduced**' transformations are very slight. Yet all three cases fall short compared to stemming or lemmatizing, indicating that for data-sparse situations, classifying all words regardless of their class is a more effective strategy.

'**Full verb**' classification achieves a 1.5 AER reduction, basically thanks to an important recall increase in all alignment directions, due to the grouping effect of this classification, so that all words belonging to a verb form become linked to the same tokens. Finally, **reordering** experiments produce very slight improvements, and apparently the result is equal no matter if the reordering is *a priori* forced as in '**Spa N-A reord**' or learnt from data as in '**N-A swap realign**'.

Combining full verb classification and stemming (of the words outside verb forms) we obtain the best AER results.

Results with the full parallel corpus are shown in Table 6. Interestingly, conclusions regarding base forms and stems do not hold in this case. Whereas base forms are not useful anymore and even degrade alignment quality, stems still provide significant improvement in AER. This is expressed in a 2.5 point absolute precision increase at a cost of 0.6 recall decrease. One possible reason for this is the harder classification of stems, especially for English, where initial vocabulary of 95K words is reduced to 81K with base forms and only 69K for stems (in Spanish, from baseline 138K vocabulary we end up with 78K base forms and 79K stems). Apparently, this involves a sparseness reduction, which makes word alignment more robust to non-literal translations. On the other hand, frequent words such as auxiliary verbs are not mapped to the same stem, thus possibly helping the aligner to discriminate compared to the case with base forms.

Partial transformations such as '**Spa lem+redPOS**', '**Spa Adj base**' and '**Spa V reduced**' do not help improve alignment quality anymore. On the other hand, '**full verb**' classification is still producing significant improvements, again reflected in the best recall figures for all alignment directions. This recall can countermeasure the recall loss when stemming and achieves the best AER (16.42) when combining these two approaches.

As about word order modification experiments, again results are not encouraging, and in this case they are even harmful for alignment quality. This holds both for deterministic Noun-Adjective reordering ('**Spa N-A reord**') and for reordering according to an initial word alignment. All combinations of order

modification and stemming, base form or verb forms classification that have been tested did not yield improvements and are not reported.

These experiments provide different alignment sets which can contain complementary information, so alignment quality can be further improved if they are combined. For the large data task, the best 3, 4 and 5 best union sets were combined with a consensus criterion. For each link present in at least one of the sets, if this link is present in a majority of sets, then it is selected for the combined set. Otherwise it is absent from the combined set. For the combination of an even number of sets, the criterion can be strict (more than half of the sets must agree) or weak (a half is enough). Results are shown in table 7. While all combinations improve the best AER presented in table 6 (that of the verbs+stems experiment), the combination of best 3 sets is particularly interesting since both recall and precision are also improved. In the 4 sets combinations, the weak criterion gives a high recall and lower precision combination, whereas the strict criterion gives a high precision but lower recall combination.

Table 7. Combination, with a consensus criterion, of the best union alignment sets obtained in the large data task (in order: the verbs+stems, stems, full verbs, spa adj base and baseline sets).

	R_S	P_P	AER
3 best	78.50	90.04	15.79
4 best (weak)	80.29	87.35	16.10
4 best (strict)	76.51	92.59	15.87
5 best	78.37	89.70	16.07

4 Conclusion and Further Work

In this paper we have evaluated the impact of performing a wide range of morphology-based data transformations in automatic word alignment. Remarkably, and even though quality improvements due to morphological information are bigger in case of data scarceness, alignment error rate can be reduced by using these informations even in case large amounts of data are available.

Specifically, stemming and verb forms classification achieve significantly better recall and precision figures in all situations. In addition, consensus combination strategies of the best alignment sets produce a further improvement of both recall and precision.

As future work, we plan to evaluate the impact of these improvements in training statistical machine translation models, as well as to define alternative translation models that incorporate useful morphological information. Additionally, other language pairs should be experimented with, as long as analysis tools and human references are available.

Acknowledgements

The authors want to thank Marta R. Costa-jussà for her invaluable help. This work has been partly supported by the TC-STAR project (European Community, FP6-506738), by Generalitat de Catalunya and by the European Social Fund.

Table 8. Some English and Spanish corpus transformations as described in corresponding sections.

	Asian countries have followed our example too .	Los países asiáticos han seguido también nuestro ejemplo .
2.1	Asian country have follow our example too .	El país asiático haber seguir también nuestro ejemplo .
2.1	asian countri have follow our exampl too .	los país asiatic han segu también nuestr ejempl .
2.1	Asian countries have followed our example too .	Los países asiático han seguido también nuestro ejemplo .
2.1	Asian countries have followed our example too .	Los países asiáticos haber#P seguido también nuestro ejemplo .
2.1	Asian countries have followed our example too .	el país_NP asiático haber_VIP3P0 seguir_VP00SM también nuestro ejemplo_NS .
2.1	Asian countries V[follow] our example too.	Los países asiáticos V[seguir] también nuestro ejemplo .
2.2	Asian countries have followed our example too .	Los asiáticos países han seguido también nuestro ejemplo .

References

- Smadja, F.A., McKeown, K.R., Hatzivassiloglou, V.: Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics* **22** (1996) 1–38
- Diab, M., Resnik, P.: An unsupervised method for word sense tagging using parallel corpora. In: *Proc. of the Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA (2002) 255–262
- Yarowsky, D., Ngai, G., Wicentowski, R.: Inducing multilingual text analysis tools via robust projection across aligned corpora. In: *Proc. of the 1st International Conference on Human Language Technology Research (HLT)*. (2001) 161–168
- Kuhn, J.: Experiments in parallel-text based grammar induction. In: *Proc. of the 42th Annual Meeting of the Association for Computational Linguistics*, Barcelona, Spain (2004) 470–477
- Brown, P., Della Pietra, S., Della Pietra, V., Mercer, R.: The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* **19** (1993) 263–311
- Zens, R., Och, F., Ney, H.: Phrase-based statistical machine translation. In *Verlag, S., ed.: Proc. German Conference on Artificial Intelligence (KI)*. (2002)

7. Mario, J., Banchs, R., Crego, J.M., de Gispert, A., Lambert, P., Fonollosa, J., Ruiz, M.: Bilingual n-gram statistical machine translation. In: Proc. of Machine Translation Summit X, Phuket, Thailand (2005) 275–82
8. Och, F., Ney, H.: A systematic comparison of various statistical alignment models. *Computational Linguistics* **29** (2003) 19–51
9. Yamada, K., Knight, K.: A syntax-based statistical translation model. In: Proc. of the Annual Meeting of the Association for Computational Linguistics, Toulouse, France (2001)
10. Och, F., Ney, H.: A comparison of alignment models for statistical machine translation. In: Proc. of the 18th Int. Conf. on Computational Linguistics, Saarbrücken, Germany (2000) 1086–1090
11. Toutanova, K., Ilhan, H.T., Manning, C.D.: Extensions to hmm-based statistical word alignment models. In: Proc. of the Conference on Empirical Methods in Natural Language Processing, Philadelphia, PA (2002)
12. Tiedemann, J.: Combining clues for word alignment. In: Proc. of the 10th Conf. of the European Chapter of the ACL (EACL), Budapest, Hungary (2003)
13. de Gispert, A.: Phrase linguistic classification and generalization for improving statistical machine translation. Proc. of the ACL Student Research Workshop (2005) 67–72
14. Popović, M., Ney, H.: Improving word alignment quality using morpho-syntactic information. In: Proc. of the 20th Int. Conf. on Computational Linguistics, COLING'04, Geneva, Switzerland (2004) 310–314
15. Popović, M., Ney, H.: POS-based word reorderings for statistical machine translation. In: Proc. 5th Int. Conf. on Language Resources and Evaluation (LREC), Genoa, Italy (2006) 1278–1283
16. Costa-jussà, M., Crego, J., de Gispert, A., Lambert, P., Khalilov, M., Banchs, R., Mariño, J., Fonollosa, J.: Talp phrase-based statistical translation system for european language pairs. In: Proc. of the HLT/NAACL Workshop on Statistical Machine Translation, New York (2006)
17. Brants, T.: Tnt — a statistical part-of-speech tagger. In: Proc. of Applied Natural Language Processing (ANLP), Seattle, WA (2000)
18. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K., Teng, R.: Five papers on wordnet. Special Issue of *International Journal of Lexicography* **3** (1991) 235–312
19. Carreras, X., Chao, I., Padr, L., Padr, M.: Freeling: An open-source suite of language analyzers. In: Proc. of the 4th Int. Conf. on Linguistic Resources and Evaluation (LREC), Lisbon, Portugal (2004)
20. Lambert, P., de Gispert, A., Banchs, R., Mario, J.: Guidelines for word alignment and manual alignment. *Language Resources and Evaluation* (2006) DOI: 10.1007/s10579-005-4822-5.
21. Och, F.: Giza++: Training of statistical translation models. <http://www.fjoch.com/GIZA++.html> (2000)