Grouping Multi-word Expressions According to Part-Of-Speech in Statistical Machine Translation

Patrik Lambert
TALP Research Center
Jordi Girona Salgado, 1-3
08034 Barcelona, Spain
lambert@gps.tsc.upc.edu

Rafael Banchs
TALP Research Center
Jordi Girona Salgado, 1-3
08034 Barcelona, Spain
rbanchs@gps.tsc.upc.edu

Abstract

This paper studies a strategy for identifying and using multi-word expressions in Statistical Machine Translation. The performance of the proposed strategy for various types of multi-word expressions (like nouns or verbs) is evaluated in terms of alignment quality as well as translation accuracy. Evaluations are performed by using real-life data, namely the European Parliament corpus. Results from translation tasks from English-to-Spanish and from Spanish-to-English are presented and discussed.

1 Introduction

Statistical machine translation (SMT) was originally focused on word to word translation and was based on the noisy channel approach (Brown et al., 1993). Present SMT systems have evolved from the original ones in such a way that mainly differ from them in two issues: first, word-based translation models have been replaced by phrase-based translation models (Zens et al., 2002) and (Koehn et al., 2003); and second, the noisy channel approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented (Och and Ney, 2002).

Nevertheless, it is interesting to call the attention about one important fact. Despite the change from a word-based to a phrase-based translation approach, word to word approaches for inferring alignment models from bilingual data (Vogel et al., 1996; Och and Ney, 2003) continue to be widely used.

On the other hand, from observing bilingual data sets, it becomes evident that in some cases it is just impossible to perform a word to word alignment between two phrases that are translations of each other. For example, certain combination of words might convey a meaning which is somehow independent from the words it contains. This is the case of bilingual pairs such as “fire engine” and “camión de bomberos”.

Notice that a word-to-word alignment strategy would most probably[1] provide the following Viterbi alignments for words contained in the previous example: “camión:truck”, “bomberos:firefighters”, “fuego:fire”, and “máquina:engine”.

Of course, it cannot be concluded from these examples that a SMT system which uses a word to word alignment strategy will not be able to handle properly the kind of word expression described above. This is because there are other models and feature functions involved which can actually *help* the SMT system to get the right translation.

However these ideas motivate for exploring alternatives for using multi-word expression information in order to improve alignment quality and consequently translation accuracy. In this sense, our idea of a multi-word expression (hereafter MWE) refers in principle to word sequences which cannot be translated literally word-to-word. However, the automatic technique studied in this work for extracting and identifying MWEs does not necessarily follow this definition rigorously.

In a preliminary study (Lambert and Banchs, 2005), we presented a technique for extracting bilingual multi-word expressions (BMWE) from parallel corpora. In that study, BMWEs identified in a small corpus[2] were grouped as a unique to-

[1] Of course, alignment results strongly depends on corpus statistics.

[2] VERBMOBIL (Arranz et al., 2003)

ken before training alignment models. As a result, both alignment quality and translation accuracy were slightly improved.

In this paper we applied the same BMWE extraction technique, with various improvements, to a large corpus (EPPS, described in section 4.1). Since this is a statistical technique, and frequencies of multi-word expressions are low (Baldwin and Villavicencio, 2002), the size of the corpus is an important factor. A few very basic rules based on part-of-speech have also been added to filter out noisy entries in the dictionary. Finally, BMWEs have been classified into three categories (nouns, verbs and others). In addition to the impact of the whole set, the impact of each category has been evaluated separately.

The technique will be explained in section 3, after presenting the baseline translation system used (section 2). Experimental results are presented in section 4. Finally some conclusions are presented and further work in this area is depicted.

2 Baseline Translation System

This section describes the SMT approach that was used in this work. A more detailed description of the presented translation system is available in Mariño et al. (2005). This approach implements a translation model which is based on bilingual n-grams, and was developed by de Gispert and Mariño (2002).

The bilingual n-gram translation model actually constitutes a language model of bilingual units which are referred to as tuples. This model approximates the joint probability between source and target languages by using 3-grams as it is described in the following equation:

$$p(T,S) \approx \prod_{n=1}^{N} p((t,s)_n | (t,s)_{n-2}, (t,s)_{n-1}) \quad (1)$$

where t refers to target, s to source and $(t,s)_n$ to the n^{th} tuple of a given bilingual sentence pair.

Tuples are extracted from a word-to-word aligned corpus. More specifically, word-to-word alignments are performed in both directions, source-to-target and target-to-source, by using GIZA++ (Och and Ney, 2003), and tuples are extracted from the union set of alignments according to the following constraints (de Gispert and Mariño, 2004):

- a monotonous segmentation of each bilingual sentence pairs is produced,
- no word inside the tuple is aligned to words outside the tuple, and
- no smaller tuples can be extracted without violating the previous constraints.

As a consequence of these constraints, only one segmentation is possible for a given sentence pair. Figure 1 presents a simple example illustrating the tuple extraction process.

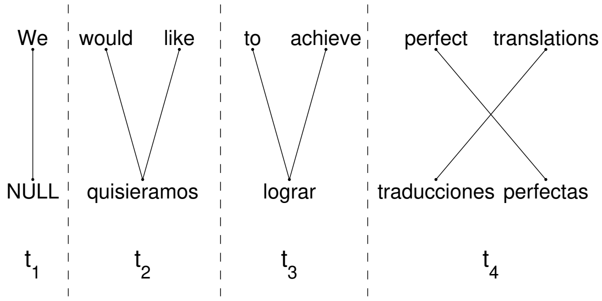

Figure 1: Example of tuple extraction from an aligned bilingual sentence pair.

A tuple set is extracted for each translation direction, Spanish-to-English and English-to-Spanish. Then the tuple 3-gram models are trained by using the SRI Language Modelling toolkit (Stolcke, 2002).

The search engine for this translation system was developed by Crego et al. (2005). It implements a beam-search strategy based on dynamic programming. The decoder's monotonic search modality was used.

This decoder was designed to take into account various different models simultaneously, so translation hypotheses are evaluated by considering a log-linear combination of feature functions. These feature functions are the translation model, a target language model, a word bonus model, a lexical model and an inverse lexical model.

3 Experimental Procedure

In this section we describe the technique used to see the effect of multi-words information on the translation model described in section 2.

3.1 Bilingual Multi-words Extraction

First, BMWEs were automatically extracted from the parallel training corpus and the most relevant ones were stored in a dictionary.

3.1.1 Asymmetry Based Extraction

For BMWE extraction, the method proposed by Lambert and Castell (2004) was used. This

	I	'm	sorry	¡	this	is	true
verdad	.	.	.	.	.	.	+
es	.	.	.	.	.	+	.
esto	.	.	.	.	+	.	.
¡	.	.	.	+	.	.	.
siento	.	–	+	.	.	.	
lo	–	ǀ	.	.	.	.	

Figure 2: There is an asymmetry in the word-to-word alignments of the idiomatic expression "lo siento – I 'm sorry". Source-target and target-source links are represented respectively by horizontal and vertical dashes.

method is based on word-to-word alignments which are different in the source-target and target-source directions, such as the alignments trained to extract tuples (section 2). Multi-words like idiomatic expressions or collocations can typically not be aligned word-to-word, and cause a (source-target and target-source) asymmetry in the alignment matrix. An asymmetry in the alignment matrix is a sub-matrix where source-target and target-source links are different. If a word is part of an asymmetry, all words linked to it are also part of this asymmetry. An example is depicted in figure 2.

In this method, asymmetries in the training corpus are detected and stored as possible BMWEs.

Accurate statistics are needed to score each BMWE entry. In the identification phase (section 3.3), these scores permit to prioritise for the selection of some entries with respect to others. Previous experiments (Lambert and Banchs, 2005) have shown than the large set of bilingual phrases described in the following section provides better statistics than the set of asymmetry-based BMWEs.

3.1.2 Scoring Based on Bilingual Phrases

Here we refer to *Bilingual Phrase* (BP) as the bilingual phrases used by Och and Ney (2004). The BP are pairs of word groups which are supposed to be the translation of each other. The set of BP is consistent with the alignment and consists of all phrase pairs in which all words within the target language are only aligned to the words of the source language and vice versa. At least one word of the target language phrase has to be aligned with at least one word of the source language phrase. Finally, the algorithm takes into account possibly unaligned words at the boundaries of the target or source language phrases.

We extracted all BP of length up to four words, with the algorithm described by Och and Ney. Then we estimated the phrase translation probability distribution by relative frequency:

$$p(t|s) = \frac{N(t,s)}{N(s)} \tag{2}$$

In equation 2, s and t stand for the source and target side of the BP, respectively. $N(t,s)$ is the number of times the phrase s is translated by t, and $N(s)$ is the number of times s occurs in the corpus. We took the minimum of both direct and inverse relative frequencies as probability of a BP. If this minimum was below some threshold, the BP was pruned. Otherwise, this probability was multiplied by the number of occurrences $N(t,s)$ of this phrase pair in the whole corpus. A weight λ was introduced to balance the respective importance of relative frequency and number of occurrences, as shown in equation 3:

$$\begin{aligned} score &= \min(p(t|s), p(s|t))\ N(t,s)^{\lambda} \\ &= \min(\frac{N(t,s)^{1+\lambda}}{N(s)}, \frac{N(t,s)^{1+\lambda}}{N(t)}) \end{aligned} \tag{3}$$

We performed the intersection between the entire BP set and the entire asymmetry based multi-words set, keeping BP scores. Notice that the entire set of BP is not adequate for our dictionary because BP are extracted from all parts of the alignment (and not in asymmetries only), so most BP are not BMWEs but word sequences that can be decomposed and translated word to word.

3.2 Lexical and Morpho-syntactic Filters

In English and Spanish, a list of stop words[3] (respectively 19 and 26) was established. The BMWE dictionary was also processed by a Part-Of-Speech (POS) tagger and eight rules were written to filter out noisy entries. These rules depend on the tag set used. Examples of criteria to reject a BMWE include:

- Its source or target side only contains stop words
- Its source or target side ends with a coordination conjunction

[3]frequently occurring, semantically insignificant words like "in", "of", "on".

- Its source or target side begins with a coordination conjunction (except "nor", in English)
- Its source or target side ends with an indefinite determiner

English data have been POS-tagged using the *TnT* tagger (Brants, 2000), after the lemmas have been extracted with *wnmorph*, included in the Wordnet package (Miller et al., 1991). POS-tagging for Spanish has been performed using the *FreeLing* analysis tool (Carreras et al., 2004).

Finally, the BMWE set has been divided in three subsets, according to the following criteria, applied in this order:

- If source AND target sides of a BMWE contain at least a verb, it is assigned to the "verb" class.
- If source AND target sides of a BMWE contain at least a noun, it is assigned to the "noun" class.
- Otherwise, it is assigned to the "misc" class (miscellaneous). Note that this class is mainly composed of adverbial phrases.

3.3 Multi-Words Identification

Identification consists, first, of the detection of all possible BMWE(s) in the corpus, and second, of the selection of the relevant candidates.

The detection part simply means matching the entries of the dictionaries described in the previous subsections. In the example of figure 2, the following BMWEs would have been detected (the number on the right is the score):

```
i am sorry ||| lo siento ||| 1566
am sorry ||| siento ||| 890
it is ||| es ||| 1004407
it is ||| esto es ||| 269
true ||| es verdad ||| 63
```

Then, selection in a sentence pair runs as follows. First, the BMWE with highest score among the possible candidates is considered and its corresponding positions are set as covered. If this BMWE satisfies the selection criterion, the corresponding words in the source and target sentences are grouped as a unique token. This process is repeated until all word positions are covered in the sentence pair, or until no BMWE matches the positions remaining to cover.

The selection criterion rejects candidates whose words are linked to exactly one word. Thus in the example, "esto – this is" would not be selected. This is correct, because the subject "esto" (this) of the verb "es" (is) in Spanish is not omitted, so that "this is – es" does not act as BMWE ("esto" should be translated to "this" and "is" to "es").

At the end of the identification process the sentence pair of figure 2 would be the following: "lo_siento ; esto es verdad – I_'m_sorry , this is true".

In order to increase the recall, BMWE detection was insensitive to the case of the first letter of each multi-word. The detection engine also allows a search based on lemmas. Two strategies are possible. In the first one, search is first carried out with full forms, so that lemmas are resorted to only if no match is found with full forms. In the second strategy, only lemmas are considered.

3.4 Re-alignment

The modified training corpus, with identified BMWEs grouped in a unique "super-token" was aligned again in the same way as explained in section 2. By grouping multi-words, we increased the size of the vocabulary and thus the sparseness of data. However, we expect that if the meaning of the multi-words expressions we grouped is effectively different from the meaning of the words they contain, the individual word probabilities should be improved.

After re-aligning, we unjoined the super-tokens that had been grouped in the previous stage, correcting the alignment set accordingly. More precisely, if two super-tokens A and B were linked together, after ungrouping them into various tokens, every word of A was linked to every word of B. Translation units were extracted from this corrected alignment, with the unjoined sentence pairs (*i.e.*the same as in the baseline). So the only difference with respect to the baseline lied in the alignment, and thus in the distribution of translation units and in lexical model probabilities.

4 Experimental Results

4.1 Training and Test Data

Our task was word alignment and translation of parliamentary session transcriptions of the European Parliament (EPPS). These data are currently available at the Parliament's website.[4] They were distributed through the TC-STAR consortium.[5] The training and translation test data used

[4]http://www.euro parl.eu.int/
[5]http: //www.tc-star.org/

included session transcriptions from April 1996 until September 2004, and from November 15th until November 18th, 2004, respectively. Translation test data include two reference sets. Alignment test data was a subset of the training data (Lambert et al., 2006).

Table 1 presents some statistics of the various data sets for each considered language: English (eng) and Spanish (spa). More specifically, the statistics presented in Table 1 are, the total number of sentences, the total number of words, the vocabulary size (or total number of distinct words) and the average number of words per sentence.

1.a.- Training data set

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1.22 M	33.4 M	105 k	27.3
Spa	1.22 M	35.0 M	151 k	28.6

1.b.- Test data set for translation

Lang.	Sentences	Words	Vocab.	Aver.
Eng	1094	26.8 k	3.9 k	24.5
Spa	840	22.7 k	4.0 k	27.0

1.c.- Word alignment reference

Lang.	Sentences	Words	Vocab.	Aver.
Eng	400	11.7 k	2.7 k	29.1
Spa	400	12.3 k	3.1 k	30.4

Table 1: Basic statistics for the considered training (a) translation test (b) and alignment test (c) data sets (M and k stands for millions and thousands, respectively).

4.2 Evaluation measures

Details about alignment evaluation can be found in Lambert et al. (2006). 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:

$$AER = 1 - \frac{|\mathcal{A} \cap \mathcal{G}_S| + |\mathcal{A} \cap \mathcal{G}|}{|\mathcal{A}| + |\mathcal{G}_S|} \quad (4)$$

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.

As for translation evaluation, we used the following measures:

WER (word error rate) or mWER (multi-reference word error rate) The WER is the minimum number of substitution, insertion and deletion operations that must be performed to convert the generated sentence into the reference target sentence. For the mWER, a whole set of reference translations is used. In this case, for each translation hypothesis, the edit distance to the most similar sentence is calculated.

BLEU score This score measures the precision of unigrams, bigrams, trigrams, and fourgrams with respect to a whole set of reference translations, and with a penalty for too short sentences (Papineni et al., 2001). BLEU measures accuracy, thus larger scores are better.

4.3 Multi-words in Training Data

In this section we describe the results of the BMWE extraction and detection techniques applied to the training data.

4.3.1 Description of the BMWE dictionaries

Parameters of the extraction process have been optimised with the alignment development corpus available with the alignment test corpus. With these parameters, a dictionary of 60k entries was extracted. After applying the lexical and morpho-syntactic filters, 45k entries were left. The best 30k entries (hereinafter referred to as *all*) have been selected for the experiments and divided in the three groups mentioned in section 3.2. *verb*, *noun* and *misc* (miscellaneous) dictionaries contained respectively 11797, 9709 and 8494 entries.

Table 2 shows recall and precision for the BMWEs identified with each dictionary. The first line is the evaluation of the MWEs obtained with the best 30k entries of the dictionary before filtering. Alignments evaluated in table 2 contained only links corresponding to the identified BMWEs. For an identified BMWE, a link was introduced between each word of the source side and each word of the target side. Nevertheless, the test data contained the whole set of links.

From table 2 we see the dramatic effect of the filters. The precision for nouns is lower than for

	Recall	Precision
Best 30k (no filters)	13.6	53.6
Best 30k (filters)	11.4	79.3
VERB (filters)	3.7	81.8
NOUN (filters)	4.0	72.8
MISC (filters)	4.1	80.8

Table 2: Quality of the BMWEs identified from the various dictionaries.

the other categories because many word groups which were identified, like "European Parliament - Parlamento europeo", are not aligned as a group in the alignment reference. Notice also that the data in table 2 reflects the precision of bilingual MWE, which is a lower bound of the precision of "super-tokens" formed in each sentence, the quantity that matters in our experiment.

Identification of BMWE based on lemmas has also been experimented. However, with lemmas, the selection phase is more delicate. With our basic selection criterion (see section 3.3), the quality of MWEs identified was worse so we based identification on full forms.

Figure 3 shows the first 10 entries in the *misc* dictionary, along with their renormalised score. Notice that "the EU - la UE", "young people - jóvenes" and "the WTO - la OMC" have been incorrectly classified due to POS-tagging errors.

```
the EU ||| la UE ||| 770731
secondly ||| en segundo lugar ||| 610599
however ||| sin embargo ||| 443042
finally ||| por último ||| 421879
firstly ||| en primer lugar ||| 324396
thirdly ||| en tercer lugar ||| 286924
young people ||| jóvenes ||| 178571
the WTO ||| la OMC ||| 174496
once again ||| una vez más ||| 169317
once ||| una vez ||| 150139
```

Figure 3: Examples of BMWEs of the *misc* category.

4.3.2 BMWE Identification Statistics

Table 3 shows, for each language, the MWE vocabulary size after the identification process, and how many times a MWE has been grouped as a unique token (instances). The different number of instances between Spanish and English correspond to one-to-many BMWEs. In general more MWEs are grouped in the Spanish side, because English is a denser language. However, the omission of the subject in Spanish causes the inverse situation for verbs.

	Vocabulary		Instances	
	ENG	SPA	ENG	SPA
ALL	12.2k	12.6k	1.28M	1.56M
VERB	6.0k	3.3k	738k	237k
NOUN	3.9k	5.9k	288k	827k
MISC	3.1k	4.3k	336k	557k

Table 3: Statistics for the BMWEs identified from the various dictionaries. ALL refers to the 30k best entries with filters.

4.4 Alignment and Translation Results

Tables 4 and 5 show the effect of aligning the corpus when the various categories of multi-words have been previously grouped.

IBM1 lexical probabilities	baseline	All
p(in_other_words|es_decir)	-	0.94
p(words|decir)	0.23	0.0013
p(other|decir)	0.026	$6\ 10^{-5}$
p(say|decir)	0.45	0.49

Table 4: Single word lexical probabilities of the alignment model in the baseline and after grouping MWE with all dictionary entries. The multi-word tokens "in_other_words" and "es_decir" do not exist in the baseline.

In table 4 we see how word-to-word lexical probabilities of the alignment model can be favourably modified. In the baseline, due to presence of the fixed expression "in other words - es decir", the probability of "words" given "decir" ("say" in English) is high. With this expression grouped, probabilities p(words|decir) and p(other|decir) vanish, while p(say|decir) is reinforced. These observations allowed to expect that with many individual probabilities improved, a global improvement of the alignment would occur.

However, table 5 shows that alignment is not better when trained with BMWEs grouped as a unique token.

A closer insight into alignments confirms that they have not been improved globally. Changes with respect to the baseline are very localised and correspond directly to the grouping of the BMWEs present in each sentence pair.

Table 6 presents the automatic translation eval-

	Recall	Precision	AER
Baseline	76.3	85.0	19.4
All	78.0	82.0	19.9
Verb	77.0	84.5	19.3
Noun	76.8	83.0	20.0
Misc	77.0	84.1	19.4

Table 5: Alignment results

uation results. In the Spanish to English direction, BMWEs seem to have a negative influence. In the English to Spanish direction, no significant improvement or worsening is observed.

	S→E		E→S	
	mWER	BLEU	mWER	BLEU
Baseline	**34.4**	**0.547**	**40.2**	**0.472**
All	36.4	0.517	40.7	0.470
Verb	35.1	0.537	**40.2**	**0.472**
Noun	35.1	0.537	40.7	0.469
Misc	35.8	0.527	41.1	0.466

Table 6: Translation results in Spanish-to-English (S→E) and English-to-Spanish (E→S) directions.

In order to understand these results better, we performed a manual error analysis for the first 50 sentences of the test corpus. We analysed, for the experiment with all dictionary entries ("All" line of table 6), the changes in translation with respect to the baseline. We counted how many changes had a neutral, positive or negative effect on translation quality. Results are shown in table 7. Notice that approximatively half of these changes were directly related to the presence some BMWE.

This study permitted to see interesting qualitative features. First, BMWEs have a clear influence on translation, sometimes positive and sometimes negative, with a balance which appears to be null in this experiment. In many examples BMWEs allowed a group translation instead of an incorrect word to word literal translation. For instance, "Red Crescent" was translated by "Media Luna Roja" instead of "Cruz Luna" (cross moon).

Two main types of error were observed. The first ones are related to the quality of BMWEs. Determiners, or particles like "of", which are present in BMWEs are mistakenly inserted in the translations. Some errors are caused by inadequate BMWEs. For example "looking at – si analizamos" ("if we analyse") cannot be used in the sense of looking with the eyes. The second type of error is related to the rigidity and data sparseness introduced in the bilingual n-gram model. For example, when inflected forms are encapsulated in a BMWE, the model looses flexibility to translate the correct inflection. Another typical error is caused by the use of back-off (n-1)-grams in the bilingual language model, when the n-gram is not any more available because of increased data sparseness.

The error analysis did not give explanation for why the effect of BMWEs is so different for different translation directions. A possible hypothesis would be that BMWEs help in translating from a denser language. However, in this case, verbs would be expected to help relatively more in the Spanish to English direction, since there are more verb group instances in the English side.

	Neutral	Positive	Negative
S→E	43	20	22
E→S	49	19	17

Table 7: Effect on quality of differences in the translations between the baseline and the BMWE experiment with "ALL" dictionary. S and E stand for Spanish and English, respectively.

5 Conclusions and Further work

We applied a technique for extracting and using BMWEs in Statistical Machine Translation. This technique is based on grouping BMWEs before performing statistical alignment. On a large corpus with real-life data, this technique failed to clearly improve alignment quality or translation accuracy.

After performing a detailed error analysis, we believe that when the considered MWEs are fixed expressions, grouping them before training helps for their correct translation in test. However, grouping MWEs which could in fact be translated word to word, doesn't help and introduces unnecessary rigidity and data sparseness in the models.

The main strength of the n-gram translation model (its history capability) is reduced when tuples become longer. So we plan to run this experiment with a phrase-based translation model. Since these models use unigrams, they are more flexible and less sensitive to data sparseness.

Some errors were also caused by noise in the automatic generation of BMWEs. Thus filtering techniques should be improved, and different methods for extracting and identifying MWEs must be developed and evaluated. Resources build manually, like Wordnet multi-word expressions, should also be considered.

The proposed method considers the bilingual multi-words as units ; the use of each side of the BMWEs as independent monolingual multi-words must be considered and evaluated.

Acknowledgements

This work has been partially funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech to Speech Translation -(IST-2002-FP6-506738, http://www.tc-star.org).

References

V. Arranz, N. Castell, and J. Giménez. 2003. Development of language resources for speech-to-speech translation. In *Proc. of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgary, September, 10-12.

T. Baldwin and A. Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Computational Natural Language Learning Workshop (CoNLL)*.

T. Brants. 2000. Tnt — a statistical part-of-speech tagger. In *Proc. of Applied Natural Language Processing (ANLP)*, Seattle, WA.

P. Brown, S. Della Pietra, V. Della Pietra, and R. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Xavier Carreras, I. Chao, L. Padró, and M. Padró. 2004. Freeling: An open-source suite of language analyzers. In *Proc. of the 4th International Conference on Linguistic Resources and Evaluation (LREC)*, Lisbon, Portugal, May.

J. M. Crego, J. Mariño, and A. de Gispert. 2005. A ngram-based statistical machine translation decoder. In *Proc. of the 9th European Conf. on Speech Communication and Technology (Interspeech)*, pages 3185–88, Lisbon, Portugal.

A. de Gispert and J. Mariño. 2002. Using X-grams for speech-to-speech translation. *Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP'02*, September.

A. de Gispert and J. Mariño. 2004. Talp: Xgram-based spoken language translation system. *Proc. of the Int. Workshop on Spoken Language Translation, IWSLT'04*, pages 85–90, October.

P. Koehn, F.J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proc. of the 41th Annual Meeting of the Association for Computational Linguistics*.

P. Lambert and R. Banchs. 2005. Data inferred multi-word expressions for statistical machine translation. In *Proc. of Machine Translation Summit X*, pages 396–403, Phuket, Tailandia.

P. Lambert and N. Castell. 2004. Alignment of parallel corpora exploiting asymmetrically aligned phrases. In *Proc. of the LREC 2004 Workshop on the Amazing Utility of Parallel and Comparable Corpora*, Lisbon, Portugal, May 25.

P. Lambert, A. de Gispert, R. Banchs, and J. Mariño. 2006. Guidelines for word alignment and manual alignment. *Accepted for publication in Language Resources and Evaluation*.

J. Mariño, R. Banchs, J. M. Crego, A. de Gispert, P. Lambert, J.A. Fonollosa, and M. Ruiz. 2005. Bilingual n-gram statistical machine translation. In *Proc. of Machine Translation Summit X*, pages 275–82, Phuket, Thailand.

G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. Miller, and R. Tengi. 1991. Five papers on wordnet. *Special Issue of International Journal of Lexicography*, 3(4):235–312.

F.J. Och and H. Ney. 2002. Dicriminative training and maximum entropy models for statistical machine translation. In *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 295–302, Philadelphia, PA, July.

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

F.J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449, December.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. IBM Research Report, RC22176, September.

A. Stolcke. 2002. SRILM: an extensible language modeling toolkit. In *Proc. of the Int. Conf. on Spoken Language Processing*, pages 901–904, Denver, CO.

S. Vogel, H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *COLING'96: The 16thInt. Conf. on Computational Linguistics*, pages 836–841, Copenhagen, Denmark, August.

R. Zens, F.J. Och, and H. Ney. 2002. Phrase-based statistical machine translation. In Springer Verlag, editor, *Proc. German Conference on Artificial Intelligence (KI)*, september.