Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation

John Tinsley, Mary Hearne, and Andy Way
Dublin City University
National Centre for Language Technology

Abstract

We use existing tools to automatically build two parallel treebanks from existing parallel corpora. We then show that combining the data extracted from both the treebanks and the corpora into a single translation model can improve the translation quality in a baseline phrase-based statistical machine translation system.

1 Introduction

The vast majority of current approaches to Machine Translation (MT) are corpus-driven. Amongst these, Phrase-Based Statistical MT (PBSMT) is by far the most dominant paradigm. Efforts to incorporate syntactic components into such systems have proven difficult. However, great strides have been taken more recently in the form of [Chiang, 2005], [Marcu et al., 2006], and [Hassan et al., 2007] amongst others. Despite these, no work to our knowledge has exploited the type of bilingual information encoded in parallel treebanks.

A parallel treebank comprises syntactically annotated aligned sentences in two or more languages. In addition to this, the trees are aligned on a sub-sentential level. In this paper we detail how we make use of freely available monolingual parsers and a statistical tree-to-tree aligner to automatically build two parallel treebanks from existing parallel corpora.

We use the (unannotated) parallel corpus to train a PBSMT system, and then investigate whether improvements can be made in translation quality by augmenting this trained system with the syntactically motivated word and phrase alignments inherent in the parallel treebanks. Finally, we look at how the alignments extracted from the parallel treebanks compare to the alignments produced using PBSMT techniques.

The remainder of this paper is organised as follows. Section 2 gives some background on phrase-based SMT and parallel treebanking. In Section 3 we describe the data we use and our experimental setup. Section 4 presents the

results of our investigation, and in Sections 5 and 6 we study the results in detail and conclude.

2 Background and Related Work

2.1 Phrase-Based Statistical Machine Translation

Phrase-based models of SMT, derived from the earlier word-based models of [Brown et al., 1990], have significantly improved the quality of statistical machine translation systems [Koehn et al., 2003]. In such models, phrases are extracted from parallel corpora using heuristics which learn phrase pairs based on word alignment data. These word alignments are extracted bidirectionally, for source→target and target→source. The heuristics then typically expand upon the intersection and union of the two sets of word alignments to learn additional phrase alignments. This is described further in Section 4.5 of [Koehn et al., 2003].

[Koehn et al., 2003] carry out experiments to investigate whether the presence of non-syntactically motivated phrases impinges on the translation quality. To do this, they discard all phrases which are not syntactically motivated, i.e. they only retain those phrases which are dominated by a single constituent in the parse tree. Their results show that this degrades translation quality.

This differs from our experiments in that the phrases we extract from the parallel treebank are by no means just a subset of the phrase alignments extractable using the PBSMT methods. Rather than restricting that set of phrase alignments to syntactically motivated phrases, we aim to augment it with all the phrase alignments we extract from the parallel treebank. We are not attempting to develop a syntax-driven model of SMT. Instead we add syntactically motivated word and phrase pairs extracted using a distinct learning technique to the PBSMT translation model in an effort show that such data can improve translation quality.

2.2 Parallel Treebanking

A parallel treebank comprises syntactically annotated aligned sentences in two or more languages. In addition to this, the sentences are aligned below the level of the clause [Volk and Samuelsson, 2004], i.e. there are alignments between nodes in the treebank which represent links between words and phrases. A simple example of a parallel treebank entry is illustrated in Figure 1.

The significance of the links encoded in parallel treebanks is not to be understated. To link two nodes is to imply translational equivalence between the surface strings dominated by both nodes. That is to say in Figure 1, the link between the source NP1 node and the target NP2 node indicates

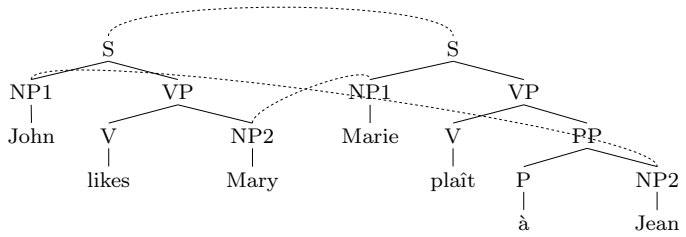

Figure 1: A typical example of an aligned English–French sentence pair in a parallel treebank

translational equivalence between the strings *John* and *Jean*. This implication is critical, because it is this alignment information which is the most useful for phrase-based SMT, as we will demonstrate in this paper.

3 Experimental Setup

3.1 Data Resources

The Parallel Corpora

In the experiments we present here, two distinct data sets were used. Firstly, from the English–Spanish section of the EuroParl corpus [Koehn, 2005], we extracted 4,911 random sentences. The only restriction was that the English sentences were required to be between 5 and 30 words in length. This set was then randomly split into 4,411 training and 500 test sentences.

The second data set consisted of 10,000 sentences extracted randomly from the English–German section of the EuroParl corpus. The restriction applied here required both English and German sentences to be between 5 and 30 words in length. This set was again randomly split into 9,000 training and 1,000 test sentences.

The Parallel Treebanks

The process of obtaining parallel treebanks from the parallel corpora described above was completely automated. Firstly, each monolingual corpus was parsed using an off-the-shelf parser. The English corpus in both data sets was parsed using Bikel's parser [Bikel, 2002] trained on the Penn II Treebank [Marcus et al., 1994]. The Spanish corpus was parsed using the same parser trained for Spanish on the Cast3LB Spanish Treebank [Civit and Martí, 2004] as described in [Chrupała and van Genabith, 2006]. Finally, the German corpus was parsed using the BitPar parser [Schmid, 2004] which was trained on the German TIGER treebank [Brants et al., 2002].

The final step in the annotation process was to automatically align the newly parsed parallel corpora at sub-sentential level. This is done by in-

serting links between constituent node pairs in the tree, which imply translational equivalence between the surface strings the nodes dominate. Tree alignment in this case is precision-based, i.e. the goal is not to aggressively align as many nodes as possible. To leave a particular node unaligned is not to say that is has no translational equivalent. Instead, an alignment higher up in the tree may encapsulate the string dominated by this node as part of a larger phrase. For example, looking back to Figure 1, even though there is no link from the source tree V node, dominating *likes*, to the target tree, does not mean it does not have a translation within this sentence pair. Instead, its translational equivalence to the non-constituent *plaît à* is captured implicitly by the links between the S and NP nodes.

To insert these sub-sentential alignments we used our own statistical tree-to-tree aligner [Tinsley et al., 2007]. This aligner induces word and phrase alignments in the parallel treebank based on word-alignment probabilities obtained from the same data data using the statistical word alignment tool, GIZA++ [Och and Ney, 2003]. The tree aligner uses these probabilities to score all hypothetical constituent pairs for a given sentence pair. Using a greedy search, it first selects the highest scoring alignments between constituents spanning phrases (both constituents must span more than one word), eliminating ill-formed links. Once completed, the process is repeated to align constituents spanning single words while respecting the constraints imposed by the higher level alignments. Thus the induced word alignments are motivated by the phrase alignments and the syntax encoded in the tree. In this way they differ from the word alignments extracted from the parallel corpus using the PBSMT method, even though they are also based on the GIZA++ tool.[1]

Given that our parallel treebanks have been automatically constructed, the question arises as to how accurate they are. Unfortunately the syntax errors that exist are unavoidable and could only be rectified by manual checking of the trees. However, the idea is that we can automatically construct parallel treebanks quite expeditely in comparison to "real" parallel treebanks, which would be impractical on the scale we are working. Also, the fact that current statistical parsers and our statistical aligner obtain high precision scores assures us that, despite errors that may exist, these parallel treebanks are still a very useful resource.

3.2 Phrase Extraction

As previously stated, we employ two principal ways of carrying out word and phrase extraction. We use the Moses [Koehn et al., 2007] system to extract words and phrases from the parallel corpus using methods mentioned in

[1]To reiterate, we induce alignments in the parallel treebank based on the GIZA++ word-alignment probability model. The PBSMT method extracts alignments from the parallel corpus based around the actual sentence-level word-alignments given by GIZA++

$\langle W_C, P_C \rangle$	Translation model consisting only of word and phrase pairs extracted from the parallel corpus using PBSMT techniques.
$\langle W_T, P_T \rangle$	Translation model consisting only of word and phrase pairs extracted from the parallel treebank based on the induced links.
$\langle W_C, P_T \rangle$	Translation model consisting of word pairs extracted from the parallel corpus and phrase pairs extracted from the parallel treebank.
$\langle W_T, P_C \rangle$	Translation model consisting of word pairs extracted from the parallel treebank and phrase pairs extracted from the parallel corpus.
$\langle W_C + W_T, P_C + P_T \rangle$	Translation model consisting of word and phrase pairs extracted from both the parallel corpus and the parallel treebank.

Table 1: Descriptions of the 5 configurations of the MT system

Section 2.1. Word and phrase pairs are extracted from the parallel treebank based on the links induced by the tree-to-tree aligner in Section 3.1. This involves extracting the string pairs dominated by each linked constituent pair in the treebank as a word or phrase alignment. Combinations of these word and phrase pairs are then used to create the translation models used in the various configurations of the MT system described in the following section.

3.3 MT System Setup

For the purposes of our experiment we create 5 different translation models using various combinations of the word and phrase pairs: $\langle W_C, P_C \rangle$, $\langle W_T, P_T \rangle$, $\langle W_C, P_T \rangle$, $\langle W_T, P_C \rangle$, and $\langle W_C + W_T, P_C + P_T \rangle$. These are described further in Table 1. For each configuration, we use Moses to estimate the word and phrase pair probabilities based on the relative frequency of the data. We also use Moses to perform the decoding task. Language modelling was carried out using the SRI language modelling toolkit [Stolcke, 2002].

Each of the five configurations of the MT system were run on both data sets in both translation directions, totalling 20 separate runs. The results of the translation output were evaluated using three automatic evaluation metrics, BLEU [Papineni et al., 2002], NIST [Doddington, 2002], and Meteor [Banerjee and Lavie, 2005].

4 Results

Looking firstly to the larger data set, the results for the English–German language pair are given in Tables 2 and 3. The configuration $\langle W_C + W_T, P_C + P_T \rangle$ improves over the baseline $\langle W_C, P_C \rangle$ across all three evaluation metrics, giving a 1% absolute increase (2.6% relative) in Meteor score from English to German, and a 1.5% increase (3.5% relative) from German to English. We now observe the results for the English–Spanish language pair in Tables 4 and 5. Again $\langle W_C + W_T, P_C + P_T \rangle$ improves over $\langle W_C, P_C \rangle$ across all three metrics, with a 1.33% alsolute increase (2.95% relative) in Meteor score from English to Spanish, and a 1.86% absolute increase (3.87% relative) from Spanish to English. These gains are all statistically significant, according to bootstrap resampling, with a confidence value of $p = 0.02$.

Looking now to the other scores in Tables 2–5, we can see that the inclusion of phrase alignments from the parallel treebank, in place of the phrase alignments extracted using PBSMT methods, has a negative effect on performance in relation to the $\langle W_C, P_C \rangle$ baseline for both data sets, in all four translation directions and across all three evaluation metrics. Conversely, inclusion of the word alignments from the parallel treebank, in place of the PBSMT word alignments, consistently improves upon the $\langle W_C, P_C \rangle$ baseline. Again this holds for both data sets in all four translation directions and across all evaluation metrics. These results and associated trends are discussed further in Section 5.

5 Discussion

The aim of these experiments was to answer the question as to whether data extracted from the parallel treebank could impact positively on translation results when combined with data extracted using PBSMT methods from a parallel corpus, within a PBSMT system framework. The results in Section 4 show this clearly to be the case. These conclusions throw up further questions regarding the respective impacts of the word and phrase pairs extracted from the parallel treebank. To examine this further, we consider the following: (i) how do the word alignments extracted from the parallel treebank compare to those extracted by a dedicated word alignment system? (ii) how do the phrases extracted by the state-of-the-art PBSMT system compare to the syntactically motivated phrases extracted from the parallel treebank?

5.1 Word Alignments

We can directly compare the PBSMT word alignments and the treebank word alignments by comparing $\langle W_T, P_T \rangle$ vs. $\langle W_C, P_T \rangle$, and $\langle W_C, P_C \rangle$ vs. $\langle W_T, P_C \rangle$ configuration pairs. Looking first to the English–German pair,

EuroParl: English ⟶ German			
Configuration	**BLEU**	**NIST**	**METEOR**
$\langle W_C, P_C \rangle$	0.1186	4.1168	0.3840
$\langle W_T, P_T \rangle$	0.1055	4.1153	0.3796
$\langle W_C, P_T \rangle$	0.1019	3.9778	0.3691
$\langle W_T, P_C \rangle$	0.1242	4.2605	0.3931
$\langle W_C + W_T, P_C + P_T \rangle$	**0.1259**	**4.3044**	**0.3938**

Table 2: English ⟶ German translation scores for various system configurations

EuroParl: German ⟶ English			
Configuration	**BLEU**	**NIST**	**METEOR**
$\langle W_C, P_C \rangle$	0.1622	4.9949	0.4344
$\langle W_T, P_T \rangle$	0.1498	5.1720	0.4327
$\langle W_C, P_T \rangle$	0.1443	4.9342	0.4176
$\langle W_T, P_C \rangle$	0.1676	5.2324	0.4473
$\langle W_C + W_T, P_C + P_T \rangle$	**0.1687**	**5.2474**	**0.4492**

Table 3: German ⟶ English translation scores for various system configurations

EuroParl: English ⟶ Spanish			
Configuration	**BLEU**	**NIST**	**METEOR**
$\langle W_C, P_C \rangle$	0.1765	4.8857	0.4515
$\langle W_T, P_T \rangle$	0.1689	4.8662	0.4560
$\langle W_C, P_T \rangle$	0.1634	4.6964	0.4440
$\langle W_T, P_C \rangle$	0.1807	5.0389	0.4619
$\langle W_C + W_T, P_C + P_T \rangle$	**0.1867**	**5.0898**	**0.4701**

Table 4: English ⟶ Spanish translation scores for various system configurations

EuroParl: Spanish ⟶ English			
Configuration	**BLEU**	**NIST**	**METEOR**
$\langle W_C, P_C \rangle$	0.1754	4.7582	0.4802
$\langle W_T, P_T \rangle$	0.1708	4.8664	0.4659
$\langle W_C, P_T \rangle$	0.1626	4.6606	0.4498
$\langle W_T, P_C \rangle$	0.1840	4.9557	0.4910
$\langle W_C + W_T, P_C + P_T \rangle$	**0.1880**	**4.9923**	**0.4935**

Table 5: Spanish ⟶ English translation scores for various system configurations

		#Tokens	#Types	COMP
EN-DE	Corpus	69,200	7,672	1,929
	Treebank	79,675	18,286	12,545
EN-ES	Corpus	37,339	5,056	904
	Treebank	43,312	11,274	7,131

Table 6: Frequency information regarding the **word** pairs extracted from the parallel corpora and treebanks, where COMP is the number of types in the relative complement of *dataset Y* in *dataset X*

as mentioned in Section 4, use of treebank word alignments leads to improvements across the board. For example, in Table 2, $\langle W_T, P_T \rangle$ improves on $\langle W_C, P_T \rangle$ by 1% Meteor score. Also, in Table 3, $\langle W_T, P_C \rangle$ sees a 1.3% increase in Meteor score over $\langle W_C, P_C \rangle$. These gains are statistically significant ($p = 0.02$).

In total, the parallel treebank provided 79,675 word pairs tokens, which yielded 18,286 unique word pair types. PBSMT extraction yielded 69,200 word pairs tokens, including 7,672 unique word pair types. 12,545 of the word pairs types extracted from the parallel treebank did not occur in the PBSMT set (the relative complement of the *PBSMT set* in the *parallel treebank set*), whereas only 1,929 of the PBSMT word pairs types did not occur in the set extracted from the parallel treebank. These figures are summarised in Table 6. Accordingly, while the parallel treebank did not provide many more word pairs tokens in total than the PBSMT extraction, it did have much broader word coverage (>50%) due to a greater variety in word pair types extracted, which in turn led to the increase in translation scores.

A similar situation can be seen with the smaller English–Spanish data set. The parallel treebank provided 43,312 word pairs tokens, yielding 11,274 unique word pair types. This can be compared to the 37,339 word pair tokens extracted using PBSMT methods, which yielded 5,056 unique types. Again there were more than twice as many word pair types extracted from the parallel treebank. Of these types, 7,131 were unique to the parallel treebank's set compared to the 904 word pair types occurring only in the PBSMT phrase extraction set. These figures are again summarised in Table 6.

These similar ratios to the English–German data led to similar translation results. In Table 4, $\langle W_T, P_T \rangle$ improves Meteor score by 1.6% over $\langle W_C, P_T \rangle$, and Table 5 shows a 1% increase in Meteor score by $\langle W_T, P_C \rangle$ over $\langle W_C, P_C \rangle$. Again, these gains are statistically significant ($p = 0.02$).

There are some indications, however, that this increase in performance may not only be attributable to the greater number of word pairs in $\langle W_T \rangle$, but also to their slightly higher quality. From the English–German data

set of 9,000 training sentence pairs, only 7,672 unique word pair types were extracted using PBSMT methods. This gives an average of less than 1 unique word alignment per sentence pair. The number of unique word pair types extracted from the parallel treebank gives an average of more than 2 unique word alignments per sentence pair. Upon looking at the ten most frequent word alignments (20 individual words) extracted from the parallel corpus only, we noted that 85% of the individual words were function words or punctuation, as opposed to 20% of those extracted from the parallel treebank only. This information seems to indicate a lot more repetition in the word alignments obtained from the parallel corpus data. However, where repetition is useful from a probability estimation perspective, the repetition here is of less informative closed class words and hence may not be as useful. Conversely, there was less repetition using the larger vocabulary of the word pairs extracted from the parallel treebank.

5.2 Phrase Alignments

As with the word alignments, we can also directly compare the phrase alignments extracted from the parallel treebank to those extracted using PBSMT methods. In this case we compare the scores of the $\langle W_T, P_T \rangle$ vs. $\langle W_T, P_C \rangle$ and $\langle W_C, P_C \rangle$ vs. $\langle W_C, P_T \rangle$ configuration pairs. Firstly, looking at the English–German language pair in Table 2, we see that the addition of the parallel treebank phrase pairs, in place of the PBSMT phrase pairs, results in a drop of 1.5% Meteor score in $\langle W_C, P_T \rangle$ over PBSMT. Similarly for German to English in Table 3, $\langle W_T, P_T \rangle$ scores 1.46% lower in Meteor score than $\langle W_T, P_C \rangle$.

In stark contrast to the numbers in Section 5.1, PBSMT phrase extraction yields more than 3 times the number of phrase pair types than are extracted from the parallel treebank. Where the parallel treebank provides 33,789 phrase pair tokens yielding 30,251 unique phrase pair types, PBSMT methods extract 120,410 phrase pair tokens yielding 97,167 unique types. This was to be expected, however, as the maximum number of possible phrases extractable from the parallel treebank is severely restricted by the syntactic structure encoded in the treebank. What was telling about these numbers is that the greater number of phrase pairs extracted using PBSMT methods did not lead to a proportionate increase in translation scores. This indicates that the phrase pairs extracted from the parallel treebank are of higher quality, but that the lack of coverage prevents the use of these phrase alignments from yielding improved translation scores on their own. These figures are summarised in Table 7 along with the statistics for the English–Spanish phrase pairs.

This assertion is given further weight by a manual inspection of the extracted phrase pairs. Figure 2 shows the ten most frequently occurring phrase pairs extracted using PBSMT methods and not from the parallel

		#Tokens	#Types	COMP
EN-DE	Corpus	120,410	97,167	92,084
	Treebank	33,789	30,251	25,041
EN-ES	Corpus	86,640	72,583	67,378
	Treebank	18,301	15,199	9,949

Table 7: Frequency information regarding the **phrase** pairs extracted from the parallel corpora and treebanks, where COMP is the number of types in the relative complement of *dataset Y* in *dataset X*

of the	⇔	der	398
that	⇔	, daß	383
president ,	⇔	präsident ,	163
mr president ,	⇔	herr präsident ,	163
european union	⇔	europäischen union	135
of the	⇔	des	127
in the	⇔	in der	122
and	⇔	, und	116
that the	⇔	daßdie	110
would like	⇔	möchte	109

Figure 2: 10 most frequently occurring phrase pairs extracted using PBSMT methods only and their frequency counts

treebank.

All but four of those phrase pairs consist solely of function words, and the average monolingual phrase length is 1.85 words. In addition, these phrases have relatively high frequency counts, i.e. they are short, frequently occurring phrase pairs which may not be very informative.

In comparison, the 10 most frequently occurring phrase pairs extracted from the parallel treebank and not from the parallel corpus in Figure 3 are more useful. This becomes apparent when we see that all but one of these phrase pairs contain at least one content word, indeed almost half of the entire vocabulary consists of content words. Also the average monolingual phrase length is 2.2 words, and these phrase pairs are generally more informative than their counterparts extracted using PBSMT methods (bar some missing German articles).

As is probably apparent by now, the English–Spanish data set is congruent with the tendencies of the English–German data set. There are almost 5 times as many parallel corpus phrase pair types than parallel treebank phrase pair types, yet the results in Tables 4 and 5 indicate that the drop in translation accuracy is even less than that for English–German. Furthermore, an almost identical situation arises with the 10 most frequent phrase pairs from the two sources.

the next item	⇔	nach der tagesordnung	22
very much	⇔	dank	17
the union 's	⇔	der union	17
the european union 's	⇔	der europischen union	15
the sitting	⇔	die sitzung	13
the commission 's	⇔	kommission	11
the resolution	⇔	den entschlieungsantrag	10
the end	⇔	ende	10
the commission	⇔	von der kommission	10
the structural funds	⇔	strukturfonds	9

Figure 3: 10 most frequently occurring phrase pairs extracted from the parallel treebank only and their frequency counts

6 Conclusions

We observe that syntactically motivated word and phrase pairs extracted from an automatically built bilingual parallel treebank have a positive effect on translation scores in a baseline PBSMT system. Combining corpus-based and treebank-based data into a single translation model gives improved translation quality when compared to a translation model comprising corpus-based data alone.

Word alignments extracted from the parallel treebank give rise to an increase in translation scores when replacing the original PBSMT word alignments. This can be attributed to the larger number of unique word alignments extracted from the parallel treebank, which provide wider word coverage and are of slightly higher quality.

Conversely, the inclusion of phrase alignments extracted from the parallel treebank, in place of the PBSMT phrase alignments, leads to a slight decrease in translation performance. However, whereas with the word alignments a slight increase in quantity resulted in a slight increase in performance, regarding phrase alignments the reverse is found. With phrase alignment there is a considerable *decrease* in quantity, up to 66% in some cases, compared to PBSMT phrase alignments. In these instances, this large decrease in quantity only results in a small drop in performance. This indicates that the phrase alignments extracted from the parallel treebank are more informative than those extracted by the PBSMT system, thus making up for the lack of quantity. This can be attributed to the linguistic information encoded in the treebank which restricts the extractable phrase pairs to those which are syntactically permissible.

Acknowledgements

This work was generously supported by Science Foundation Ireland, Grant

No. 05/RF/CMS064. We thank Ventsislav Zhechev and Declan Groves for their assistance at various stages of the work and the anonymous reviewers for their insightful comments.

References

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-05)*, Ann Arbor, MI.

[Bikel, 2002] Bikel, D. (2002). Design of a Multi-lingual, parallel-processing statistical parsing engine. In *Human Language Technology Conference (HLT)*, San Diego, CA, USA.

[Brants et al., 2002] Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria.

[Brown et al., 1990] Brown, P. F., Cocke, J., Pietra, S. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., and Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16(2):79–85.

[Chiang, 2005] Chiang, D. (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 263–270, Ann Arbor, MI.

[Chrupała and van Genabith, 2006] Chrupała, G. and van Genabith, J. (2006). Using Machine-Learning to Assign Function Labels to Parser Output for Spanish. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 136–143, Sydney, Australia. Association for Computational Linguistics.

[Civit and Martí, 2004] Civit, M. and Martí, M. A. (2004). Building Cast3LB: A Spanish Treebank. *Research on Language and Computation*, 2(4):549–574.

[Doddington, 2002] Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In *Human Language Technology: Notebook Proceedings*, pages 128–132, San Diego, CA.

[Hassan et al., 2007] Hassan, H., Sima'an, K., and Way, A. (2007). Supertagged Phrase-based Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07)*, pages 288–295, Prague, Czech Republic.

[Koehn, 2005] Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open Source Toolkit for

Statistical Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, pages 177–180, Prague, Czech Republic.

[Koehn et al., 2003] Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-Based Translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL '03)*, pages 48–54, Edmonton, Canada.

[Marcu et al., 2006] Marcu, D., Wang, W., Echihabi, A., and Knight, K. (2006). SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-06)*, pages 44–52, Sydney, Australia.

[Marcus et al., 1994] Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 110–115, Princeton, NJ.

[Och and Ney, 2003] Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, PA.

[Schmid, 2004] Schmid, H. (2004). Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 04)*, Geneva, Switzerland.

[Stolcke, 2002] Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference Spoken Language Processing*, Denver, CO.

[Tinsley et al., 2007] Tinsley, J., Zhechev, V., Hearne, M., and Way, A. (2007). Robust Language-Pair Independent Sub-Tree Alignment. In *Machine Translation Summit XI*, pages 467–474, Copenhagen, Denmark.

[Volk and Samuelsson, 2004] Volk, M. and Samuelsson, Y. (2004). Bootstrapping Parallel Treebanks. In *Proceedings of the 7th Conference of the Workshop on Linguistically Interpreted Corpora (LINC)*, pages 71–77, Geneva, Switzerland.