TALN 2008, Avignon, 9–13 juin 2008

Comparing Constituency and Dependency Representations for SMT Phrase-Extraction

Mary Hearne, Sylwia Ozdowska and John Tinsley
National Centre for Language Technology, Dublin City University,
Glasnevin, Dublin 9, Ireland
{mhearne,sozdowska,jtinsley}@computing.dcu.ie

Résumé. Nous évaluons le recours à des techniques de traduction à base de segments syntaxiquement motivés, seules ou en combinaison avec des techniques à base de segments non motivés, et nous comparons les apports respectifs de l'analyse en constituants et de l'analyse en dépendances dans ce cadre. À partir d'un corpus parallèle Anglais–Français, nous construisons automatiquement deux corpus d'entraînement arborés, en constituants et en dépendances, alignés au niveau sous-phrastique et en extrayons des correspondances bilingues entre mots et syntagmes motivées syntaxiquement. Nous mesurons automatiquement la qualité de la traduction obtenue par un système à base de segments. Les résultats montrent que la combinaison des correspondances bilingues non motivées et motivées sur le plan syntaxique améliore la qualité de la traduction quel que soit le type d'analyse considéré. Par ailleurs, le gain en qualité est plus important avec le recours à l'analyse en dépendances au regard des constituants.

Abstract. We consider the value of replacing and/or combining string-based methods with syntax-based methods for phrase-based statistical machine translation (PB-SMT), and we also consider the relative merits of using constituency-annotated *vs.* dependency-annotated training data. We automatically derive two subtree-aligned treebanks, dependency-based and constituency-based, from a parallel English–French corpus and extract syntactically motivated word- and phrase-pairs. We automatically measure PB-SMT quality. The results show that combining string-based and syntax-based word- and phrase-pairs can improve translation quality irrespective of the type of syntactic annotation. Furthermore, using dependency annotation yields greater translation quality than constituency annotation for PB-SMT.

Mots-clés : Traduction statistique à base de segments, annotation en constituants, annotation en dépendances, corpus parallèles arborés alignés au niveau sous-phrastique.

Keywords: PB-SMT, constituency annotation, dependency annotation, subtree-aligned parallel treebanks.

1 Introduction

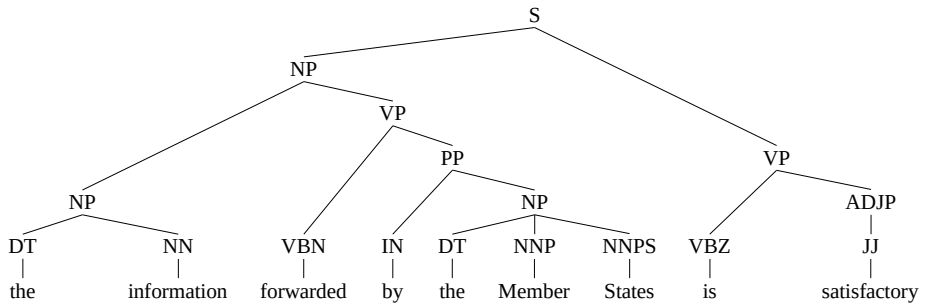
The standard technique used to induce translation models from parallel corpora (Koehn *et al.*, 2003) is not motivated by linguistic information. That is, all phrase-pairs compatible with a given word-alignment for any sentence pair are extracted; the word-alignment process is not syntax-aware and, generally, the only criterion for phrase-pair exclusion is phrase length. It seems reasonable that the incorporation of linguistic knowledge into the phrase-extraction process could yield better results, either because additional useful phrase pairs could then be identified or because it would allow for the exclusion of less useful phrase correspondences.

An experiment by Koehn *et al.* (2003) suggests that the latter hypothesis does not hold true: when phrase-pairs not corresponding to syntactic constituents were discarded, translation accuracy decreased. However, work presented by both Groves & Way (2005) and Tinsley *et al.* (2007a) suggests that the first hypothesis is valid. In (Groves & Way, 2005), phrase-pairs are extracted from sentence-aligned data by first chunking the sentences monolingually using stop-word information and then aligning those chunks using mutual information techniques together with relative chunk position information. While replacing the standard phrase-pairs with these novel string-pairs led to a reduction in translation accuracy, combining both sets of phrase alignments gave improved translation scores. (Tinsley *et al.*, 2007a) took a somewhat different approach, first constituency-parsing the training sentence pairs, then aligning node pairs using a statistical tree aligner (Tinsley *et al.*, 2007b) and finally extracting all string pairs dominated by linked constituents. Again, replacing the standard phrase-pairs with these novel tree-based pairs did not improve translation accuracy, but results increased when both sets of phrase alignments were combined. Crucially, in these approaches, the extracted data were not based on any a priori fixed word alignment. Thus, many new phrase-pairs not discovered by the original method were made available during translation, resulting in improved accuracy.

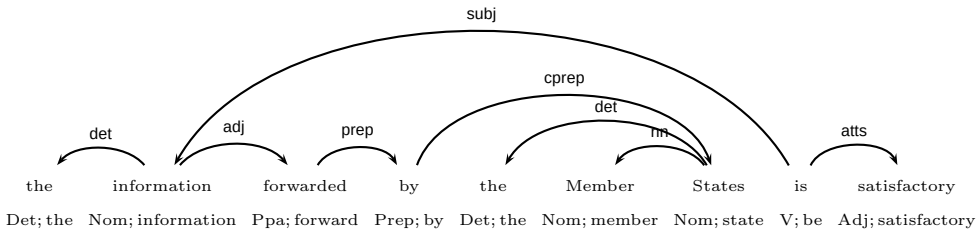
In this paper, we investigate the impact of variation in syntactic analysis type – specifically, constituency parsing *vs.* dependency parsing – on the translation model induction technique introduced in (Tinsley *et al.*, 2007a). Our experimental objective is to compare the relative value of phrase-pairs which can be extracted from each type of representation to phrase-based statistical machine translation (PB-SMT) by measuring translation accuracy. Thus, we automatically construct two subtree-aligned parallel treebanks, one dependency-parsed and the other constituency-parsed, from a single parallel corpus. We take a tree-aligner previously used only to align constituency trees and describe how we used it to align dependency trees. We induce phrase-translation models from the resulting datasets and carry out translation experiments using the Moses decoder (Koehn *et al.*, 2007). We evaluate the output using standard evaluation metrics and present our findings.

2 Annotations, Data and Tools

The data annotation types we consider in this work are constituency parses and dependency parses. In both cases, each sentence is tagged with part-of-speech information, and in the case of dependency parses a lemma is also associated with each word. Constituency parses, or context-free phrase-structure trees, make explicit syntactic constituents (such as noun phrases (NP), verb phrases (VP) and prepositional phrases (PP)) identifiable in the sentence. An example of a constituency parse is given in Fig. 1, where we see that the overall sentence comprises an NP followed by a VP, each of which has some internal structure. Dependency parses make explicit the relationships between the words in the sentence in terms of heads and dependents. An example of a dependency parse is given in Fig. 2, where an arc from word w_i to word w_j indicates that w_i is w_j's head and, correspondingly, w_j is w_i's dependent. These arcs are labelled such that the label indicates the nature of the dependency – in the given example, the label on the arc from *is* to *information* is labelled SUBJ indicating that *information* is the subject.



(1)

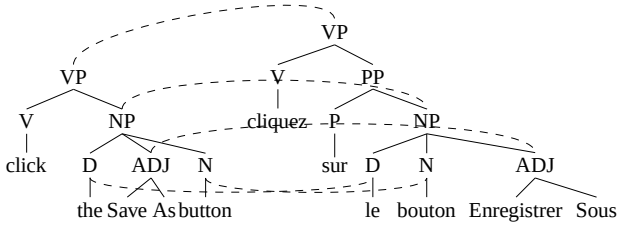


(2)

In the experiments we present here, we used the JOC English–French parallel corpus provided within the framework of the ARCADE campaigns to evaluate sentence alignment and translation spotting (Chiao *et al.*, 2006).[1] The JOC corpus is composed of texts published in 1993 as a section of the C Series of the Official Journal of the European Community. It contains about 400,000 words corresponding to 8,759 aligned sentences with an average sentence length of 23 words for English and 27.2 words for French.

[1]The JOC corpus is distributed by ELRA/ELDA (`www.elda.org`).

Two subtree-aligned parallel treebanks were automatically derived from this dataset using off-the-shelf tools (parsers and aligner). Each treebank comprises syntactic annotations – constituency-based and dependency-based – and alignments between source and target nodes which make explicit the translational equivalences between words and phrases. For dependency parsing, we used the English and the French versions of SYNTEX (Bourigault *et al.*, 2005). For constituency parsing, we used Bikel's statistical parser (Bikel, 2002) trained on the Penn II Treebank (Marcus *et al.*, 1994) for English and the Modified French Treebank (Schluter & van Genabith, 2007) for French.



(3)

Previous work has seen the development of tools which automatically induce alignments between parsed sentence pairs. Here, we use the tool described in (Tinsley *et al.*, 2007b). This tool is designed to discover an optimal set of alignments between any given, fixed tree pair, independent of language pair and constituent labelling schema. It requires a single external resource: the two word-alignment probability models output by GIZA++ (Och & Ney, 2003) when trained in both directions on parallel text for the language pair being aligned.[2]

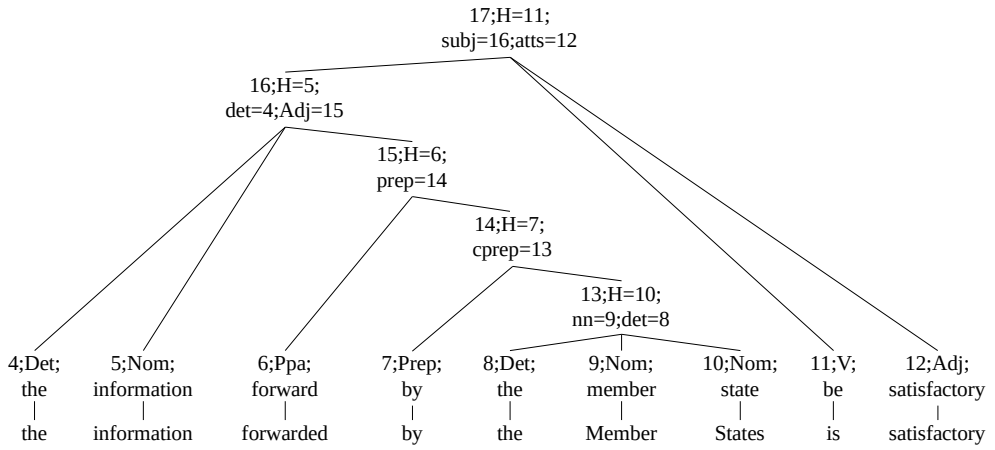
The tree-aligner works by hypothesising all possible alignments between the nodes in the tree pair. It scores each of these hypotheses using the GIZA++ word-alignment probabilities as described in detail in (Tinsley *et al.*, 2007b). Using a greedy search, it then iteratively selects the highest-scoring alignment hypothesis and eliminates all hypotheses that conflict with it. The tree-alignment is complete when no non-zero-scored, non-conflicting hypotheses remain. An example of a constituency parsed tree-aligned sentence pair is given in Fig. 3.

The tree-aligner used here has not previously been used to align dependency structures. These structures are not directly compatible with the aligner because the tool requires that the input trees be in labelled, bracketed format. While the labels themselves can be arbitrary and the branching-factor and depth of the tree are irrelevant – for instance, a part-of-speech-tagged sentence with a single, arbitrary root label is perfectly acceptable – it must be possible to associate each node in the tree with its corresponding surface string. The output of the dependency parser, as shown in Fig. 2, does not directly

[2]Minimally, this parallel text should comprise the sentence pairs from the parallel treebank being aligned, but it can, of course, be extended to include all available parallel text for the language pair in question.

meet this requirement and we must therefore convert the dependency-parsed data into a bracketed structure that the aligner can handle. Note that this conversion is formal rather than linguistic. As the aligner does not look inside node labels and our experiments require only the extraction of string-pairs from the aligner output, we pack sufficient information into the node labels such that the original dependency information is fully recoverable from the bracketed structure.

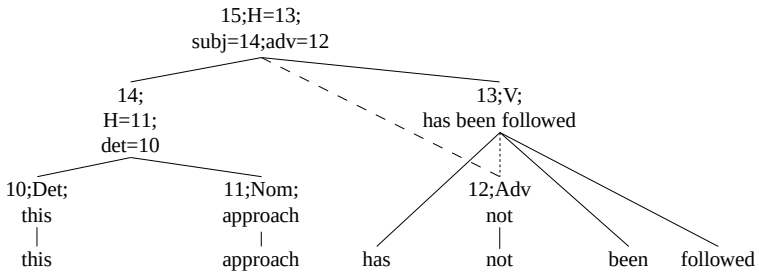
The bracketed representation for the dependency structure in Fig. 2 is given in Fig. 4. In this representation, each constituent is comprised of a head and its dependents arranged as siblings in the order in which they occurred in the sentence. Each node label retains the dependency information, indicating which child is head and the function of each of its dependent children. The label formats for constituents and parts-of-speech are *index;head=index;func*$_1$*=index;...;func*$_n$*=index* and *index;tag;lemma* respectively.



(4)

The single feature of dependency parses which cannot be satisfactorily encoded in our bracketed representation is non-projectivity. An example of a non-projective dependency structure is given in Fig. 5. In our bracketed representation, each head and its direct dependents are grouped as siblings under a single node according to the surface word order. In Fig. 5, the relationship between the dependent *not* and its head *has been followed* is correctly represented by the dashed line from the root constituent 15 to constituent 12. However, as this branch crosses the one between 13 and *has*, this structure is not acceptable to the aligner. This forces us to compromise by attaching the non-projective constituent to the lowest non-crossing parent constituent. Thus, the dashed line in Fig. 5 is dropped and the dotted line linking 12 to 13 is inserted instead. However, the true relationship is encoded in the node labelling: constituent 15's label records the fact that 13 is 12's head.[3]

[3]This analysis arises from the parser's pre- and post-processing procedures, which result in deviations from standard part-of-speech tagging.



(5)

3 Experiments

As described in previous sections, we have constructed three different versions of the JOC English–French parallel corpus, the first containing aligned sentence pairs, the second containing aligned constituency tree pairs and the third containing aligned dependency structure pairs. While the dataset comprises 8759 pairs, 37 were discarded because they could not be assigned an English and/or French constituency parse. The dataset was then split into 1000 test/reference pairs and 7722 training pairs, and the same split applied to all three versions. In all experiments presented here, the source language is French and the target English.

Our experimental objective is to compare the value of phrase-pairs which can be extracted from the different dataset representations to PB-SMT by measuring translation accuracy. All translation experiments are carried out using the Moses system (Koehn *et al.*, 2007). The evaluation metrics used are BLEU (Papineni *et al.*, 2002), NIST (Doddington, 2002) and METEOR (Banerjee & Lavie, 2005). These metrics all compare each output translation to a reference translation in terms of the substrings they have in common.

3.1 Phrase-Table Computation

Phrase-pairs are extracted from the string-aligned training data by standard PB-SMT techniques using the Moses system. Each sentence-pair is first word-aligned using GIZA++ in both source-to-target and target-to-source directions. After obtaining the intersection of these directional alignments, alignments from the union are also inserted; this insertion process is heuristics-driven (Koehn *et al.*, 2003). Once the word-alignments are finalised, all word- and phrase-pairs (overlapping and non-overlapping) which are consistent with the word-alignment and which comprise 7 words or less are extracted. Frequency counts for the extracted pairs are computed over the entire training set. This dataset is henceforth referred to as STR.

Extracting phrase-pairs from the constituency-aligned and dependency-aligned datasets involves extracting the string pairs dominated by each linked node pair in the treebank as

	BLEU	NIST	METEOR
STR	30.35	62.62	64.32
CON	29.97	63.19	63.59
DEP	29.90	63.32	64.11
STR+CON	31.98	65.16	65.61
STR+DEP	32.03	65.28	65.72
STR+CON+DEP	30.97	63.40	64.75

Table 1: Evaluation of translation accuracy using the extracted phrase-pair sets both individually and in combination.

a word or phrase alignment (Tinsley *et al.*, 2007a); all word- and phrase-pairs (overlapping and non-overlapping) which are consistent with the tree-alignment are extracted. Frequency counts for the extracted pairs are computed over the entire training set. These datasets are henceforth referred to as CON and DEP.

3.2 Results

We compute a variety of final phrase-tables based on combinations of the STR, CON and DEP phrase-pair sets. In all cases, the final probabilities assigned are relative frequencies based on the frequency counts from each dataset being included. The results of our experiments are presented in Tab. 1. In analysing our results, we considered both the value of replacing and/or combining string-based methods with syntax-based methods, and also the relative merits of using constituency-annotated *vs.* dependency-annotated data for PB-SMT. Our observations are as follows:

- replacing the standard string-based phrase-extraction method with either of the syntax-based methods (STR and CON in Tab. 1) tends to result in a decrease in translation accuracy, respectively 1.78% and 1.24% relative decrease in BLEU, 1.24% and 0.5% in METEOR;
- combining the standard string-based phrase-extraction method with either of the syntax-based methods (STR+CON and STR+DEP) leads to improved translation accuracy, respectively 5.04% and 5.34% relative increase in BLEU, 3.86% and 3.91% in NIST and 1.97% and 2.15% in METEOR;
- combining all three approaches (STR+CON+DEP) does not yield greater accuracy than combining STR with just one of the syntax-based phrase-tables: over STR+CON, the relative increases are of 0.06% and 0.09% in BLEU and NIST respectively; over STR+DEP, there is a relative increase of 0.05% in NIST;
- annotating the training data with dependency structures generally yields greater translation accuracy than annotating it with constituency structures for PB-SMT:

DEP outperforms CON by 0.17% for NIST and by 0.74% for BLEU; STR+DEP outperforms STR+CON according to all metrics by between 0.05% and 0.28%.

In some instances, the evaluation metrics give conflicting results for the same system output. For example, we see that for BLEU, CON improves over DEP, but the opposite is true for NIST and METEOR. No one measure in particular is accepted as being most reliable. However, it is generally accepted that a significant increase, or decrease, in all three metrics is conclusive. A clear-cut increase in all scores can be observed only when combining the string-based method with either of the syntax-based methods. All other combinations show a general trend toward preferring the dependency annotation without the results being conclusive.

A potential explanation for this latter observation may lie in differences between those phrase-pair types which were extracted from one parallel treebank but not the other. Previous experiments have shown that shorter phrase-pairs have greater impact on translation accuracy (Koehn *et al.*, 2003). While fewer unique phrase-pair types were extracted from the dependency-annotated treebank (15,078, *vs.* 20,571 phrase-pair types which occurred in the constituency-annotated treebank only) these phrases are shorter on average (5.67 *vs.* 9.98 tokens per phrase) and may go some way towards explaining the overall preference for the dependency parses. Furthermore, there are more linked constituency-based phrases (66,601 for English and 67,280 for French *vs*. 64,904 and 64,135 respectively for the dependency-based phrases). This higher alignment coverage may lead to lower accuracy, thus having a negative impact on translation quality.

Of course, differing translation accuracies may also be due to differences between the monolingual parses generated, either because of inherent divergences between the dependency and constituency representations or because of disparities in parser accuracy. Regarding the quality of the parsers, the reported accuracies are reversed according to language: reported f-scores for the constituency parser are 90% for English (Bikel, 2002) and 80% for French (Schluter & van Genabith, 2007), whereas reported f-scores for the dependency parser are 82% for English and 89% for French (Ozdowska, 2006)[4]. However, these scores were obtained for datasets not directly comparable to the one used here. As we do not have gold-standard parses for our dataset, we cannot report parse accuracy figures, but it is nevertheless clear that the syntactic representations for each monolingual dataset differed significantly. A quantitative comparison of the English non-POS constituents output (i.e. constituents spanning more than 1 word) shows that 52% were unique to the constituency-parser output and that 38.9% were unique to the dependency-parser output; the same analysis of the French constituents shows 46.6% unique to the constituency-parser output and 42.7% unique to the dependency-parser output. We hope that further analysis will shed more light on the importance of this issue.

[4]See (Paroubek *et al.*, 2007) for a standard evaluation of parsers for French, including SYNTEX.

4 Conclusions and Future Work

We observe that PB-SMT benefits from syntactically motivated word- and phrase-pairs derived out of constituency-annotated and dependency-annotated subtree-aligned treebanks with a general trend towards preferring the dependency representation. Combining string-based extraction with either of the constituency-based or dependency-based extraction results in significantly improved translation quality over a baseline string-based extraction. However, combining all three extraction methods does not yield greater accuracy. Comparing constituency annotations *vs.* dependency annotations, we conclude that dependency-based extraction performs significantly better either individually or in conjunction with string-based extraction.

In future work, we plan to scale up these experiments and analyse the relative impact of the different types of constituents for which phrase pairs are extracted in order to gain further insights into the usefulness of syntactic information for PB-SMT. We will also look in more detail at the issue of relative phrase length.

Acknowledgements
Thanks to Natalie Schluter for providing us with the Modified French Treebank and the French language package for Bikel's parser. Thanks to Ventsislav Zhechev for his assistance with the tree aligner. Thanks also to the anonymous reviewers for their insightful comments. This work was funded by Science Foundation Ireland grants 05/RF/CMS064 and 05/IN/1732.

References

BANERJEE S. & 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)*, p. 65–72, Ann Arbor, MI.

BIKEL D. M. (2002). Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, p. 24–27, San Francisco, CA.

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M.-P. & OZDOWSKA S. (2005). SYNTEX, analyseur syntaxique de corpus. In *Atelier EASy, Actes de la Conférence Traitement Automatique des Langues Naturelles*, Dourdan, France.

CHIAO Y.-C., KRAIF O., LAURENT D., NGUYEN T. M. H., SEMMAR N., STUCK F., VÉRONIS J. & ZAGHOUANI W. (2006). Evaluation of multilingual text alignment systems: the ARCADE II project. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, p. 1975–1978, Genoa, Italy.

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

GROVES D. & WAY A. (2005). Hybrid Example-Based SMT: the Best of Both Worlds? In *Proceedings of ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, p. 183–190, Ann Arbor, MI.

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. & HERBST E. (2007). Moses: Open source toolkit for statistical machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session*, p. 177–180, Prague, Czech Republic.

KOEHN P., OCH F. J. & 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)*, p. 48–54, Edmonton, Canada.

MARCUS M., KIM G., MARCINKIEWICZ M. A., MACINTYRE R., BIES A., FERGUSON M., KATZ K. & SCHASBERGER B. (1994). The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the ARPA Workshop on Human Language Technology*, p. 110–115, Princeton, NJ.

OCH F. J. & NEY H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, **29**(1), 19–51.

OZDOWSKA S. (2006). ALIBI, *un système d'ALIgnement BIlingue à base de règles de propagation syntaxique*. Phd. Thesis, University Toulouse-Le Mirail, France.

PAPINENI K., ROUKOS S., WARD T. & 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)*, p. 311–318, Philadelphia, PA.

PAROUBEK P., VILNAT A., ROBBA I. & AYACHE C. (2007). Les résultats de la campagne EASy d'évaluation des analyseurs syntaxiques du français. In *Actes de la Conférence Traitement Automatique des Langues Naturelles*, Toulouse, France.

SCHLUTER N. & VAN GENABITH J. (2007). Preparing, Restructuring and Augmenting a French Treebank: Lexicalised Parsing or Coherent Treebanks? In *Proceedings of the 10th Conference of the Pacific Association of Computational Linguistics (PACLING 2007)*, Melbourne, Australia.

TINSLEY J., HEARNE M. & WAY A. (2007a). Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation. In *Proceedings of The Sixth International Workshop on Treebanks and Linguistic Theories (TLT-07)*, Bergen, Norway.

TINSLEY J., ZHECHEV V., HEARNE M. & WAY A. (2007b). Robust Language Pair-Independent Sub-Tree Alignment. In *Proceedings of MT Summit XI*, Copenhagen, Denmark.