CimS - The CIS and IMS Joint Submission to WMT 2015 addressing morphological and syntactic differences in English to German SMT

Fabienne Cap[1], **Marion Weller**[1,2], **Anita Ramm**[1] **and Alexander Fraser**[1]
[1] CIS, Ludwig-Maximilian University of Munich – (cap|ramm|fraser)@cis.uni-muenchen.de
[2] IMS, University of Stuttgart – wellermn@ims.uni-stuttgart.de

Abstract

We present the CimS submissions to the WMT 2015 Shared Task for the translation direction English to German. Similar to our previous submissions, all of our systems are aware of the complex nominal morphology of German. In this paper, we combine source-side reordering and target-side compound processing with basic morphological processing in order to obtain improved translation results. We also report on morphological processing for English to French.

1 Introduction

This paper presents our submissions to the WMT shared task 2015. We use customised solutions to address morphological challenges in the English to German translation direction. Our goal is to make German and English as similar as possible in order to obtain better word alignments and hence an improved translation quality. We base our work on three main components, which we have carefully investigated separately in the past.

(i) Nominal Inflection We use context-based prediction of German inflectional endings. This improves fluency and enables the creation of morphological forms which have not occurred in the training data.

(ii) Source-side Reordering We reorder the English source text in order to make it more similar to the German word order. This improves word alignment and thus translation quality. It also makes the reordering task in decoding easier.

(iii) Compound Processing We split German compounds into simple words for training. In decoding, we translate only simple words, some of which are re-combined into compounds afterwards in post-processing. This allows us to create compounds which have not occurred in the training data.

This year, our main focus is on combining nominal inflection prediction and source-side reordering. We investigated both of these components separately in the past and expect an additive positive effect on translation quality when combined. We then added compound processing, which we already have investigated in combination with nominal inflection before, but not together with source-side reordering. Here, we also expect the combination to outperform the single components in terms of translation quality.

2 Methodology

The underlying idea of all of our systems is to improve translation quality by making the source and target languages more similar than they usually are. We address three common problems in English to German SMT: morphological richness in terms of inflectional variants, productive compounding and different word orders. In Figure 1, we illustrate the latter two of these problems using an example sentence which contains both a German compound (*"Mehrheitsvotum"* = "majority vote") and different word orders.

The methods we use to solve all three of these problems are implemented as pre- and post-processing steps. For nominal inflection and compound handling, the German data is transformed into an underspecified representation prior to training. After translation we transform the underspecified output into fluent German by merging some adjacent words into compounds and generating suitable inflectional endings. As for the differing word orders of German and English, only one pre-processing step is required, reordering the English source sentences into German word order.

In this section, we describe the different steps in more detail.

Proceedings of the Tenth Workshop on Statistical Machine Translation, pages 84–91,
Lisboa, Portugal, 17-18 September 2015. ©2015 Association for Computational Linguistics.

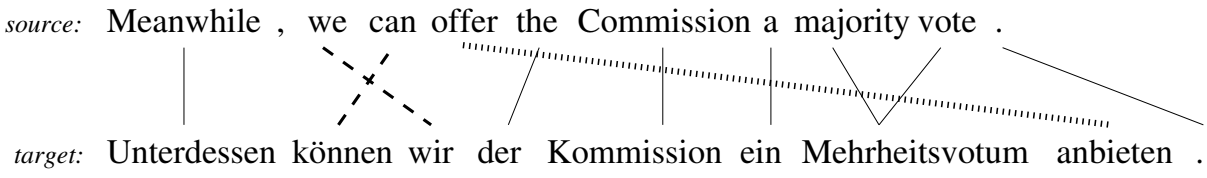

Figure 1: Illustration of structural differences between English and German. Dashed and dotted lines indicate a different word order, while the bold lines indicates a potentially problematic 1:n alignment due to a compound. Such structural differences may lead to erroneous word alignments.

stemmed SMT output with feature markup	morph. features	generated forms	gloss
auf[APPR-auf-**Dat**]	–	auf	*on*
die<+ART><Def>[ARTdef]	Fem.Dat.Sg.St	der	*the*
Tag<NN>Ordnung<+NN><**Fem**><**Sg**>[NN]	Fem.Dat.Sg.Wk	Tagesordnung	*agenda*
stehen[VVFIN]	–	stehen	*are*
die<+ART><Def>[ARTdef]	Masc.Nom.Pl.St	die	*the*
Plan<+NN><**Masc**><**Pl**>[NN]	Masc.Nom.Pl.Wk	Pläne	*plans*
für[APPR-für-**Acc**]	–	für	*for*
eine<+ART><Indef>[ARTindef]	Fem.Acc.Sg.St	eine	*a*
groß<+ADJ><Comp>[ADJA]	Fem.Acc.Sg.St	größere	*bigger*
nuklear<+ADJ><Pos>[ADJA]	Fem.Acc.Sg.St	nukleare	*nuclear*
Zusammenarbeit<+NN><**Fem**><**Sg**>[NN]	Fem.Acc.Sg.Wk	Zusammenarbeit	*co-operation*

Table 1: Overview of the morphology-aware SMT system for the input sentence "*... on the agenda are plans for greater nuclear co-operation*".

2.1 Morphology-aware SMT

In order to build an SMT system which is aware of German nominal inflection, the German data is reduced to a lemmatised representation, which contains translation-relevant morphological features (stem-markup, cf. first column in Table 1). This stem-markup consists of *number* and *gender* annotated at nouns: *gender* is considered as part of the lemma of a noun. The annotation of *number* onto target-side nouns aims at preserving the number of the source phrase during translation, as we expect nouns to be translated with their appropriate number value. This markup is only applied to nouns, i.e. the head of NPs or PPs, because the grammatical features of adjectives and determiners are dependent on the translation context in which they appear. For nominal inflection, the morphological features *number, gender, case and strong/weak inflection* need to be modelled. For each of the four morphological features, we use a linear chain CRF (Lafferty et al. (2001)) trained on stems/lemmas and the respective feature, using the Wapiti toolkit (Lavergne et al., 2010). During feature prediction, the features that are set by the stem-markup (*number, gender* on nouns) are propagated over the rest of the linguistic phrase. In contrast, *grammatical case* depends on the role of the NP in the sentence (e.g. subject or direct/indirect object) and is therefore determined entirely from the surrounding context in the sentence. The value for *strong/weak inflection* depends on the combination of the other features, cf. second column in Table 1. Based on the lemma and the predicted features, inflected forms are then generated using the rule-based morphological analyser SMOR (Schmid et al., 2004), cf. third column in Table 1.

Even though this basic nominal inflection does not handle compounds, it is able to model simple word formation processes: portmanteau prepositions (prepostion+determiner, e.g. *zum=zu+dem* "to the") are split in pre-processing and re-merged in the post-processing step, following a simple set of rules (e.g. merging only in singular, never in plural for a limited set of prepositions).

2.2 Reordering

The different word order of clauses in English and German may often lead to misaligned verbal elements. While German verbs often occur in clause-final position, English verbs mostly appear in rigid SVO order. We parsed the English section of the parallel data with (Charniak and Johnson, 2005) using a model we trained on the standard Penn Treebank sections. The scripts we used for reordering the English input are similar to the ones we previously described in (Gojun and Fraser, 2012). Figure 2 illustrates how reordering

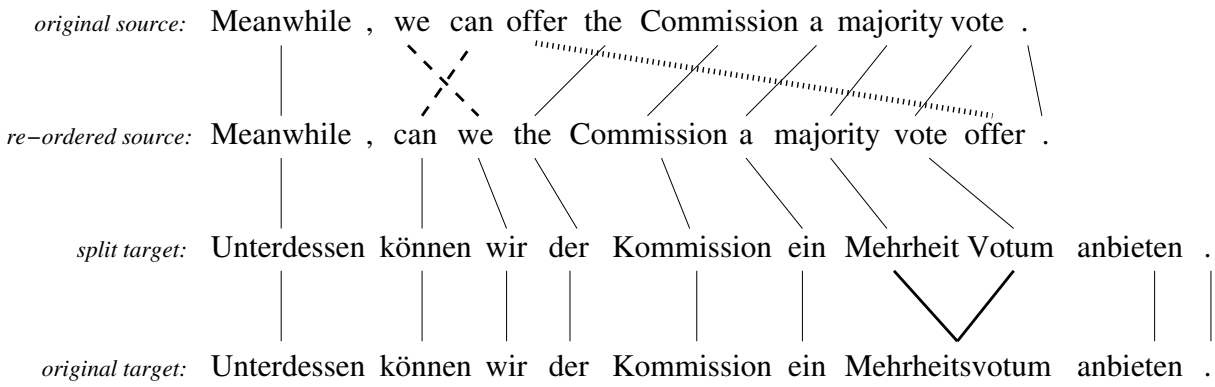

Figure 2: Illustration of how re-ordering the English input may help to reduce crossing and long-distance alignments and how target-side compound splitting may transform 1:n into 1:1 alignments.

the English input sentence can lead to less crossing and long-distance alignments.

2.3 Compound Processing

German allows for closed compounds where in English two or more words are required to express a certain content. This asymmetry can lead to alignment and thus translation errors. Moreover, German allows for **productive** compounding, i.e. new compounds can be generated from scratch and may not have occurred in the training data. Compound processing solves these two problems through splitting compounds for translation and, when translating into German, deciding whether to recombine words into compounds based on the context.

For compound splitting we use a rule-based morphological analyser where ambiguous analyses are disambiguated using corpus statistics. In general, we follow the method described in (Fritzinger and Fraser, 2010) for splitting: we disambiguate multiple analyses using context-sensitive POS and corpus-based word frequencies. The example given in Figure 2 shows how compound splitting can transform a 1:n alignment into a 1:1 alignment.

Note that for English to German translation, we always combine compound processing with nominal inflection prediction in order to maximise the generalisation over seen word parts in the training corpus. We thus translate from English into a split and underspecified version of German. Then, in a second step, compounds are merged using sequence prediction of good merge points (based on source language and target language features). Finally, words taking nominal inflection are re-inflected using the nominal inflection procedure. More details can be found in (Cap et al., 2014a).

3 Experimental Settings

For the WMT shared task, we combined the three components which we have described in the previous section. An overview of all systems we trained can be found in Table 2.

Data For all of our systems, we exclusively used data distributed for the WMT shared task 2015. We used all of the available monolingual data for German and all of the available parallel data for German and English.

UTF8 Cleaning Even though the submitted training data is provided in UTF-8 encoding, it contains a considerable number of characters that are not cleanly encoded into UTF8. We identified these characters and sequences thereof by reading all data bytewise and mapping it to the main UTF-8 encoding tables covering the Western European languages. All lines that contained one or more characters which did not fit these tables – either because they have been broken or because they belong to non-latin scripts like, e.g., Chinese or Arabic, were removed from the corpora as we expected those lines to lead to erroneous analyses in the subsequent preprocessing steps of our pipeline.

Length Constraints To ensure good alignment quality, we removed sentence pairs where one language is considerably longer than the other (pairs exceeding the ratio 1:9 words), as well as sentences containing many special characters (e.g. several dashes in row) indicating that the line in question is part of e.g. a table. Furthermore, we removed all sentences with a sentence length of more than 100 words. Table 3 gives an overview of the parallel data after cleaning and pre-processing.

Experiment	portmanteau merging	nominal inflection	source-side re-ordering	compound merging
Inflection^{Contrastive}	+	+		
Inflection_Reordering^{Primary}	+	+	+	
Inflection_Compounds	+	+		+
Inflection_Reordering_Compounds	+	+	+	+

Table 2: Names and components of our SMT systems; the submitted system are named *CIMS-primary* and *CIMS*.

	original	encoding	length or ratio	not parseable	cleaned
News	272,807	203	1,381	12,095	259,128
Europarl	1,920,209	24	17,637	3,855	1,898,693
CommonCrawl	2,399,123	17,508	7,489	26,623	2,347,503
parallel data	4,592,139	17,735	37,221	289,606	4,505,324

Table 3: Overview of the parallel data after cleaning and pre-processing.

English Variants The English source-side is mapped into British English in order to make the data as consistent as possible.

Linguistic Preprocessing The abstract representation for the nominal inflection requires the annotation of morphological features. After tokenization, we thus parsed all target-side data with BitPar (Schmid, 2004). To obtain the lemmas and suitable compound splittings, we applied SMOR (Schmid et al., 2004).

Language Model We trained 5-gram Language Models for each of the available German monolingual corpora and the German sections of the parallel data. For each corpus (the monolingual news corpora 07-14 and the parallel corpora europarl, commoncrawl and news), we built separate language models using the SRILM toolkit (Stolcke, 2002) with Kneser-Ney smoothing and then interpolated[1] them using weights optimized on development data (cf. tuning set 08-13). We then used KenLM (Heafield, 2011) for faster processing.

We performed this language model training for two different kinds of experiments: those **without** compound processing are trained on the underspecified (= lemmatised) representation, while those **with** compound processing are trained on a split underspecified representation.

Phrase-based Translation Model For word alignment, we use the multi-threaded GIZA++ toolkit (Och and Ney, 2003; Gao and Vogel, 2008). Our translation models were trained using Moses (Koehn et al., 2007), following the instructions for a baseline shared task system, using default settigs. All our systems are trained identically – what differs is the degree to which the underlying training data has been modified.

Tuning We tuned feature weights using batchmira with 'safe–hope' (Cherry and Foster, 2012) until convergence (or up to 25 runs). We used the tuning data of all previous shared tasks from 2008 to 2013, which gave us 16,071 sentences for tuning. We tuned each experiment separately against an underspecified (i.e. lemmatised) version of the tuning reference optimising BLEU scores (Papineni et al., 2002). Note also that we integrated the CRF-based compound prediction and merging procedure for each experiment with compound processing into each tuning iteration and thus scored the output against a non-split lemmatised reference.

Testing After decoding, some post-processing is required in order to retransform the underspecified representation into fluent German text. Our post-processing consists of the following steps:

1) translate into (split) underspecified German
2) merge compounds
3) predict nominal inflection
4) merge portmanteaus

Finally, the output was recapitalised and detokenised using the shared task tools and all available German training data. We calculated BLEU scores using the NIST script version 11b.

[1] /mosesdecoder/scripts/ems/support/interpolate-lm.perl

Experiment	news2014	news2015
	$BLEU_{ci}$	$BLEU_{ci}$
submitted contrastive: Inflection	–	21.46
submitted primary: Inflection_Reordering	–	21.65
Raw	19.92	21.44
Raw_Portmanteau	19.83	21.54
Inflection	19.86	21.49
Inflection_Reordering	20.35	21.64
Inflection_Compounds	19.08	20.43
Inflection_Reordering_Compounds	19.65	21.19

Table 4: BLEU scores for all our systems. The upper part lists the submitted results (using a language model built on a subset of the available data), the lower part compares all our variants which have been computed after the deadline with a language model based on all available data for the constrained task.

4 Results

For evaluation, we used the 3,003 sentences of the 2014 shared task as well as the 2,169 sentences of this year's shared task. The results are given in Table 4. In the upper part of the table we present the results for the submitted systems, in the lower part we compare all variants of our systems. Note that we compare our systems against two baselines: *Raw* denotes a system built on all parallel and monolingual data available for the shared task, while *Raw_Portmanteau* denotes a system based on the same data, though restricted to parseable sentences, as we split portmanteaus based on POS tags.

It can be seen that dealing with nominal inflection alone does not considerably improve or decrease the BLEU scores of the two baselines. However, the combination of nominal inflection and source-side reordering has a positive effect on translation quality. When it comes to the combination of compound processing and nominal inflection, which we have successfully applied in the past (Cap et al., 2014a; Cap et al., 2014b), we do not see any improvement in terms of BLEU score for this combination here. This does not necessarily mean that the compound systems quality is worse, as previous manual evaluations have shown that BLEU scores do not adequately reflect all compound-related improvements in translation quality (Cap et al., 2014a). Finally the results given in Table 4 show that adding source-side reordering to the combination of compound processing and nominal inflection does improve the BLEU scores, even though they still remain lower than for nominal inflection and source-side reordering without compound processing. We have never combined all three components before, but despite the lower performance in terms of BLEU scores we will further pursue this combination in the future.

4.1 Comparison to Other Shared Task Submissions

In addition to automatic metrics, the shared task submissions are also manually evaluated. In this evaluation, our primary system (BLEU score of 21.65) was placed in a cluster with 4 other systems, of which at least two have BLEU scores of 23 and higher. Furthermore, our system was placed in a cluster ranked higher in the manual evaluation than a cluster containing a single system with a BLEU score of 22.6 (one BLEU point higher than our system). This shows clearly that BLEU underestimates the quality of our submission. Despite its comparatively low BLEU scores it is perceived to be of similar or better quality than systems with considerably higher BLEU scores when judged by human annotators. This supports our hypothesis that morphological modeling in combination with reordering improves translation quality and is consistent with human evaluations of morphological modeling we have carried out in the past, see, e.g., (Weller et al., 2013; Cap et al., 2014a).

5 Additional Experiments: English to French translation

In an additional set of experiments, we applied the nominal inflection system also to an English–French system.

Nominal Inflection for French The general pipeline is the same as for translation into German.

We used RFTagger for French (Schmid and Laws, 2008) for morphological tagging and a French version of SMOR to generate inflected forms. The stem-markup on the French data corresponds to that of the German markup (*number* and *gender* on nouns). In contrast to four morphological features for nominal inflection in German, only *number* and *gender* need to be modelled for French.

Data The EN–FR data set is much larger than that for EN–DE; after applying the same preprocessing steps, we obtained a parallel corpus of more than 36 million sentence pairs. For the language model, we used an additional 45.9 million lines (news07-14 and newsdiscuss corpus). The language model was interpolated over separate language models built on the different corpora using the development set to obtain optimal weights.

Results The results of the submitted systems are shown in the table below:

Raw		**Nominal Inflection**^{P}	
BLEU$_{ci}$	BLEU$_{cs}$	BLEU$_{ci}$	BLEU$_{cs}$
32.24	31.19	32.26	31.22

The nominal inflection system is our primary system. Due to the large amount of EN–FR parallel training data, we assume that here the BLEU score correctly shows that there is not much difference in performance between the two systems.

6 Previous Work

Nominal Inflection The approach we use for nominal inflection prediction which was first described by (Toutanova et al., 2008). The approach consists of two steps: i) translate into an underspecified representation of German (most words being lemmatised) and ii) after translation predict inflectional endings depending on the actual context of the word(s). While developed for Russian and Arabic morphology, we adapted the approach of Toutanova et al. (2008) to the needs of German in (Fraser et al., 2012). In (Weller et al., 2013), we extended this work to use subcategorisation information and source-side syntactic features in order to improve the accuracy of case prediction. Note that we did not use this extension of our pipeline in the present shared task.

Reordering Different word orders have already been addressed in previous approaches. For example, Collins et al. (2005) reordered German prior to translating into English, which lead to improved translations. In (Gojun and Fraser, 2012), we switched the translation direction and reordered the English input sentence before translating into German, which in turn resulted in improved translation quality.

Compound Processing In the past, there have been numerous attempts to address compound splitting for German to English. Almost every German to English SMT system nowadays incorporates some kind of compound processing, either using corpus-based word frequencies (Koehn and Knight, 2003), POS-contraints (Stymne et al., 2008), lattice-based approaches (Dyer, 2009) or language-independent segmentation (Macherey et al., 2011). In our work we have been using a rule-based morphological analyser combined with corpus statistics for compound splitting (Fritzinger and Fraser, 2010), a procedure which we have updated since that work. Details can be found in (Cap et al., 2014a).

For compound merging, we translate from English into split and lemmatized German. Then, in a second step, compounds are merged using a CRF-based approach based on (Stymne and Cancedda, 2011) and then re-inflected using the nominal inflection procedure as described above. More details of our compound merging approach can be found in (Cap et al., 2014a).

7 Conclusion and Future Work

In our submission to WMT 2015, we combined the three components nominal inflection, source-side reordering and compound processing. We expected a positive effect on translation quality above the performance of each of these components when applied in isolation.

While this effect was not evident in the obtained BLEU scores, the manual evaluation, in which our system was found to be of equal or better quality than systems achieving higher BLEU scores, makes it clear that in fact our approaches do improve translation quality.

Our current systems are built on the standard version of Moses with default settings; as part of future work we plan to investigate better strategies to exploit Moses' numerous methods for optimization.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 644402 (HimL) and the DFG grants *Distributional Approaches to Semantic Relatedness* and *Models of Morphosyntax for Statistical Machine Translation (Phase 2)*.

References

Fabienne Cap, Alexander Fraser, Marion Weller, and Aoife Cahill. 2014a. How to Produce Unseen Teddy Bears: Improved Morphological Processing of Compounds in SMT. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*.

Fabienne Cap, Marion Weller, Anita Ramm, and Alexander Fraser. 2014b. CimS - The CIS and IMS joing submission to WMT 2014 – Translating from English to German. In *Proceedings of the 9th Workshop on Statistical Machine Translation at ACL, System Papers*.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL)*.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause Restructuring for Statistical Machine Translation. In *Proceedings of 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Chris Dyer. 2009. Using a Maximum Entropy Model to Build Segmentation Lattices for MT. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL HLT)*.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling Inflection and Word Formation in SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Fabienne Fritzinger and Alexander Fraser. 2010. How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In *Proceedings of the Fifth Workshop on Statistical Machine Translation*.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*.

Anita Gojun and Alexander Fraser. 2012. Determining the Placement of German Verbs in English-to-German SMT. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Association for Computational Linguistics, Demonstration Session*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML'01: Proceedings of the 18th International Conference on Machine Learning*.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical Very Large Scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Klaus Macherey, Andrew M. Dai, David Talbot, Ashok C. Popat, and Franz Och. 2011. Language-independent Compound Splitting with Morphological Operations. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics (ACL)*.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Helmut Schmid and Florian Laws. 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-grained POS Tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: a German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of The 20th International Conference on Computational Linguistics (COLING)*.

Andreas Stolcke. 2002. SRILM – an Extensible Language Modelling Toolkit. In *ICSLN'02: Proceedings of the international conference on spoken language processing*.

Sara Stymne and Nicola Cancedda. 2011. Productive Generation of Compound Words in Statistical Machine Translation. In *Proceedings of the 6th Workshop on Statistical Machine Translation and Metrics MATR of the Conference on Empirical Methods in Natural Language Processing*.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2008. Effects of Morphological Analysis in Translation between German and English. In *Proceedings of the 3rd Workshop on Statistical Machine Translation*.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Marion Weller, Alexander Fraser, and Sabine Schulte im Walde. 2013. Using Subcategorization Knowledge to Improve Case Prediction for Translation to German. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*.