CimS – The CIS and IMS joint submission to WMT 2014 translating from English into German

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

Abstract

We present the CimS submissions to the 2014 Shared Task for the language pair EN→DE. We address the major problems that arise when translating into German: complex nominal and verbal morphology, productive compounding and flexible word ordering. Our morphology-aware translation systems handle word formation issues on different levels of morpho-syntactic modeling.

1 Introduction

In our shared task submissions, we focus on the English to German translation direction: we address different levels of productivity of the German language, i.e., nominal and verbal inflection and productive word formation, which lead to data sparsity and thus confuse classical SMT systems.

Our basic goal is to make the two languages as morphosyntactically similar as possible. We use a parser and a morphological analyser to remove linguistic features from German that are not present in English and reorder the English input to make it more similar to the German sentence structure. Prior to training, all words are lemmatised and compounds are split into single words. This is not only beneficial for word alignment, but it also allows us to generalise over inflectional variants of the same lexemes and over single words which could occur in one place as a standalone word and in another place as part of a compound. Translation happens in two steps: first, we translate from English into split, lemmatised German and then, we perform compound merging and generation of inflection as a post-processing step. This way, we are able to create German compounds and inflectional variants that have not been seen in the parallel training data.

In this paper, we investigate the performance of well-established source-side reordering, nominal re-inflection and compound processing systems on an up-to-date shared task. In addition, we present experimental results on a verbal inflection component and a syntax-based variant including source-side reordering.

2 Related Work

Re-Inflection The two-step translation approach we use was described by e.g. Toutanova et al. (2008) and Jeong et al. (2010), who use a number of morphological and syntactic features derived from both source and target language. More recently, Fraser et al. (2012) describe a similar approach for German using different CRF-based feature prediction models, one for each of the four grammatical features to be predicted for German words in noun phrases, namely *number*, *gender*, *case* and *definiteness*. This approach also handles word-formation issues such as portmanteau splitting and compounding. Weller et al. (2013) added subcategorization information in combination with source-side syntactic features in order to improve the prediction of *case*.

De Gispert and Mariño (2008) generate verbal inflection for translation from English into Spanish. They use classifiers trained not only on target language but also on source language features, which is even more crucial for the prediction of verbs than it is for nominal inflection.

More recently, Williams and Koehn (2011) translate directly into target language surface forms. Agreement within NPs and PPs, and also between subject and verb is considered during the decoding process: they use string-to-tree translation, where the target language (German) morphology is expressed as a set of unification constraints automatically learned from a morphologically annotated German corpus.

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 71–78,
Baltimore, Maryland USA, June 26–27, 2014.

Compound Processing Compound splitting for SMT has been addressed by numerous different groups, for translation from German to English, e.g. using corpus-based frequencies (Koehn and Knight, 2003), using POS-constraints (Stymne et al., 2008), a lattice-based approach propagating the splitting decision to the decoder (Dyer, 2009), a rule-based morphological analyser (Fritzinger and Fraser, 2010) or unsupervised, language-independent segmentation (Macherey et al., 2011).

Compound processing in the other translation direction, however, has been much less investigated. Popović et al. (2006) describe a list-based approach, in which words are only re-combined if they have been seen as compounds in a huge corpus. However this approach is limited to the list's coverage. The approach of Stymne (2009) overcomes this coverage issue by making use of a POS-markup which distinguishes former compound modifiers from former heads and thus allows for their adequate recombination after translation. An extension of this approach is reported in Stymne and Cancedda (2011) where a CRF-model is used for compound prediction. In Cap et al. (2014) their approach is extended through using source-language features and lemmatisation, allowing for maximal generalisation over compound parts.

Source-side Reordering One major problem in English to German translation is the divergent clausal ordering: in particular, German verbs tend to occur at the very end of clauses, whereas English sticks to a rigid SVO order in most cases. Collins et al. (2005), Fraser (2009) and Gojun and Fraser (2012) showed that restructuring the source language so that it corresponds to the expected structure of the target language is helpful for SMT.

3 Inflection Prediction

German has a rich morphology, both for nominal and verbal inflection. It requires different forms of agreement, e.g., for adjectives and nouns or verbs and their subjects. Traditional phrase-based SMT systems often get such agreements wrong. In our systems, we explicitly model agreement using a two-step approach: first we translate from English into lemmatised German and then generate fully inflected forms in a second step. In this section, we describe our nominal inflection component and first experimental steps towards verbal re-inflection.

3.1 Noun Phrase Inflection

Prior to training, the German data is reduced to a lemmatised representation containing translation-relevant morphological features. For nominal inflection, the lemmas are marked with *number* and *gender*: *gender* is considered as part of the lemma, whereas *number* is indirectly determined by the source-side, as we expect nouns to be translated with their appropriate *number* value. We use a linear chain CRF (Lafferty et al., 2001) to predict the morphological features (*number*, *gender*, *case* and *strong/weak*). The features that are part of the lemma of nouns (*number*, *gender*) are propagated over the rest of the linguistic phrase. In contrast, *case* depends on the role of the NP in the sentence (e.g. subject or direct/indirect object) and is thus to be determined entirely from the respective context in the sentence. The value for *strong/weak* depends on the combination of the other features. Based on the lemma and the predicted features, inflected forms are then generated using the rule-based morphological analyser SMOR (Schmid et al., 2004). This system is described in more detail in Fraser et al. (2012).

3.2 Verbal Inflection

German verbs agree in number and person with their subjects. We thus have to derive this information from a noun phrase in nominative case (= the subject) near the verb. This information comes from the nominal inflection prediction described in section 3.1. We predict tense and mode of the verb using a maximum-entropy classifier which is trained on English and German contextual information. After deriving all information needed for the generation of the verbs, the inflected forms are generated with SMOR.

4 Compound Processing

In English to German translation, compound processing is more difficult than in the opposite direction. Not only do compounds have to be split accurately, but they also have to be put together correctly after decoding. The disfluency of MT output and the difficulty of deciding which single words should be merged into compounds make this task even more challenging.

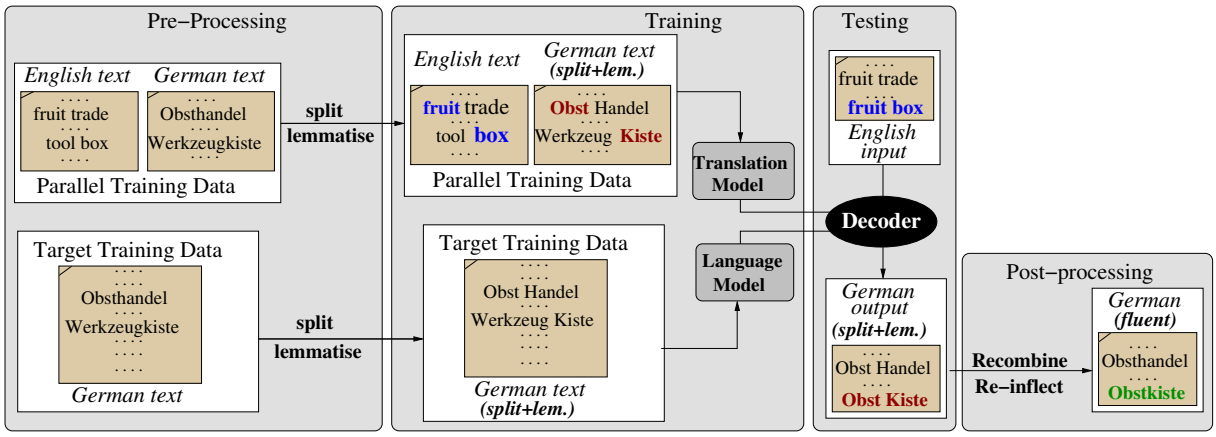

Figure 1: Pipeline overview of our primary CimS-CoRI system.

We combine compound processing with inflection prediction (see Section 3) and thus extend the two-step approach respectively: compounds are split and lemmatised simultaneously, again using SMOR. This allows for maximal generalisation over former compound parts and independently occurring simple words. We use this split representation for training. After decoding, we re-combine words into compounds again, using our extended CRF-based approach, which is based on Stymne and Cancedda (2011), but includes source-language features and allows for maximal generalisation through lemmatisation. More details can be found in Cap et al. (2014). We then use SMOR to generate sound German compounds (including morphological transformations such as introduction or deletion of filler letters). Finally, the whole text including the newly-created compounds, is re-inflected using the nominal inflection prediction models as described in Section 3.1 above. This procedure allows us to create compounds that have not been seen in the parallel training data, and also inflectional variants of seen compounds. See Figure 1 for an overview of our compound processing pipeline.

4.1 Portmanteaus

Portmanteaus are a special kind of compound. They are a fusion of a preposition and a definite article (thus not productive) and their *case* must agree with the *case* of the noun. For example, *"zum"* can be split into *"zu"* + *"dem"* = to+the$_{Dative}$. They introduce additional sparsity to the training data: imagine a noun occurred with its definite article in the training data, but not with the portemanteau required at testing time. Splitting portemanteaus allows a phrase-based SMT system to access phrases covering nouns and their corresponding definite articles. In a post-processing step, definite articles are then re-merged with their preceding prepositions to restore the original portmanteau, see (Fraser et al., 2012) for details. This generalisation effect is even larger as we not only split portmanteaus, but also lemmatise the articles.

5 System descriptions

Our shared task submissions include different combinations of the inflection and compound processing procedures as described in the previous two sections. We give an overview of all our systems in Table 1. Note that we did not re-train the compound processing CRFs on the new dataset, but used our models trained on the 2009 training data instead. However, this does not hurt performance, as the CRF we use is not trained on surface forms, but only frequencies and source-side features instead. See (Fraser et al., 2012) and (Cap et al., 2014) for more details on how we trained the respective CRFs. In contrast, the verbal classifier has been trained on WMT 2014 data.

6 Experimental Settings

In all our systems, we only used data distributed for the shared task. All available German data was morphologically analysed with SMOR. For lemmatisation of the German training data, we disambiguated SMOR using POS tags we obtained through parsing the German section of the parallel training data with BitPar (Schmid,

No.	apprart splitting	nominal inflection	compound processing	verbal inflection	source-side reordering
CimS-RI	X	X			
CimS-CoRI^P	X	X	X		
CimS-RIVe	X	X		X	
CimS-CoRIVe	X	X	X	X	
CimS-Syntax-RORI	X	X			X

Table 1: Overview of our submission systems.RI = nominal **R**e-**I**nflection, Co = **Co**mpound processing, Ve = **Ve**rbal inflection, RO = source-side **R**e-**O**rdering. Syntax = syntax-based SMT ^{P} = primary submission.

2004) and tagging the big monolingual training data using RFTagger (Schmid and Laws, 2008)[1]. Note that we did not normalise German language e.g. with respect to old vs. new writing convention etc. as we did in previous submissions (e.g. (Fraser, 2009)).

For the compound prediction CRFs using syntactic features derived from the source language, we parsed the English section of the parallel data using EGRET, a re-implementation of the Berkeley-Parser by Hui Zhang[2]. Before training our models on the English data, we normalised all occurrences of British vs. American English variants to British English. We did so for training, tuning and testing input.

Language Model We trained 5-gram language models based on all available German monolingual training data from the shared task (roughly 1.5 billion words) using the SRILM toolkit (Stolcke, 2002) with Kneser-Ney smoothing. We then used KenLM (Heafield, 2011) for faster processing. For each of our experiments, we trained a separate language model on the whole data set, corresponding to the different underspecified representations of German used in our experiments, e.g. lemmatised for *CimS-RI*, lemmatised with split compounds for *CimS-CoRI*, etc.

Phrase-based Translation model We performed word alignment using the multithreaded GIZA++ toolkit (Och and Ney, 2003; Gao and Vogel, 2008). For translation model training and decoding, we used the Moses toolkit (Koehn et al., 2007) to build phrase-based statistical machine translation systems, following the instructions for the baseline system for the shared task, using only default settings.

Syntax-based Translation model As a variant to the phrase-based systems, we applied the inflection prediction system to a string-to-tree system with GHKM extraction (Galley et al. (2004), Williams and Koehn (2012)). We used the same data-sets as for the phrase-based systems, and applied BitPar (Schmid, 2004) to obtain target-side trees. For this system, we used source-side reordering according to Gojun and Fraser (2012) relying on parses obtained with EGRET[3].

Tuning For tuning of feature weights, we used *batch-mira* with *'–safe-hope'* (Cherry and Foster, 2012) until convergence (or maximal 25 runs). We used the 3,000 sentences of *newstest2012* for tuning. Each experiment was tuned separately, optimising Bleu scores (Papineni et al., 2002) against a lemmatised version of the tuning reference. In the compound processing systems we integrated the CRF-based prediction and merging procedure into each tuning iteration and scored each output against the same unsplit and lemmatised reference as the other systems.

Testing After decoding, the underspecified representation has to be retransformed into fluent German text, i.e., compounds need to be re-combined and all words have to be re-inflected. The whole procedure can be divided into the following steps:

1a) translation into lemmatised German representation (RI, RIVe)
1b) translation into split and lemmatised German (CoRi, CoRIVe)
2) compound merging (CoRI, CoRIVe):
3) nominal inflection prediction and generation of full forms using SMOR (all)
4) verbal re-inflection (RIVe, CoRIVe)
5) merging of portmanteaus (all)

[1] We could not parse the whole monolingual dataset due to time-constraints and thus used RFTagger as a substitute.

[2] available from `https://sites.google.com/site/zhangh1982/egret`.

[3] Note that we observed some data-related issues on the Syntax-RORI experiments that we hope to resolve in the near future.

Experiment	mert.log news2012	Bleu ci news2013	Bleu cs news2013	Bleu ci news2014	Bleu cs news2014
raw	16.52	18.62	17.61	17.80	17.25
CimS-RI	18.51	19.23	18.38	18.33	17.75
CimS-CoRI^P	18.36	19.13	18.25	18.51	17.87
CimS-RIVe	19.08	18.89	18.06	17.86	17.31
CimS-CoRIVe	18.69	18.60	17.77	17.38	16.78
CimS-Syntax-RORI	18.26	19.04	18.17	18.15	17.59

Table 2: Bleu scores for all CimS-submissions of the 2014 shared task. ci = case-insensitive, cs = case-sensitive; ^{P} = primary submission.

After these post-processing steps, the text was automatically recapitalised and detokenised, using the tools provided by the shared task, which we trained on the whole German dataset. We calculated Bleu (Papineni et al., 2002) scores using the NIST script version 13a.

7 Results

We evaluated our systems with the 3,000 sentences of last year's *newstest2013* and also the 2,737 sentences of the 2014 blind test set for the German-English language pair. The Bleu scores of our systems are given in Table 2, where *raw* denotes our baseline system which we ran without any pre- or postprocessing whatsoever. Note that the big gap in mert.log scores between *raw* and the CimS-systems comes from the fact that *raw* is scored against the original (i.e. fully inflected) version of the tuning reference, while the CimS-systems are scored against the stemmed tuning reference.

As for the Bleu scores of the test sets, we observe similar improvements for the CimS-RI and CimS-CoRI systems of +0.5/0.6 with respect to the *raw* baseline as we did in previous experiments (Cap et al., 2014)[4]. In contrast, our systems incorporating verbal prediction inflection (CimS-RIVe/CoRIVe) cannot yet catch up with the performance of the well-investigated nominal inflection and compound processing systems (CimS-RI/CoRI). We attribute this partly to the positive influence we assume fully inflected verbs to have in nominal inflection prediction models, but as the verb processing systems are still under development, there might be other issues we have not discovered yet. We plan to revisit these systems and improve them.

Finally, the syntax-based reordering system yields scores that are competitive to those of CimS-RI/CoRI. While Syntax-RORI so far only incorporates source-side reordering and nominal re-inflection, we plan to investigate further extensions of this approach in the future.

7.1 Additional Evaluation

We manually screened the filtered 2014 test set and identified 3,456 German compound tokens, whereof 862 did not occur in the parallel training data and thereof, 244 did not even occur in the monolingual training data. For each of our systems, we calculated the number of compound reference matches they produced. The results are given in Table 3.

system	ref	new
raw	827	0
CimS-RI .	864	5
CimS-CoRI^P	1,064	109
CimS-RIVe	853	5
CimS-CoRIVe	1,070	122
CimS-Syntax-RORI	900	20

Table 3: Numbers of compounds produced by the systems that matched the reference (*ref*) and did not occur in the parallel training data (*new*).

The compound processing systems (with Co in the name) generate many more correct compounds than comparable systems without compound handling. Compared to the raw baseline, CoRI/CoRIVe did not only produce 237/243 more reference matches, but also 109/122 compounds that matched the reference but did not occur in the parallel training data. A lookup of those 109/122 compounds in the monolingual training data (consisting of roughly 1.5 billion words) revealed, that 8/6 of them did not oc-

[4]We will have a closer look at the data from a compound processing view in Section 7.1 below.

cur there either[5]. These were thus not accessible to a list-based compound merging approach either. This result also shows that despite the fact that CoRIVe does not yield a competitive translation quality performance yet, the compound processing component seems to benefit from the verbal inflection and it is definitely worth more investigation in the future.

Moreover, it can be seen from Table 3 that the re-inflection systems (*RI*) produce more reference matches than the raw baseline. Interestingly, they even produce some reference matches that have not been seen in the parallel training data due to inflectional variation, and in the case of the syntax-based system due to a naive list-based compound merging: even though it has not been trained on a split representation of German text, it might occasionally occur that two German nouns occur next to each other in the MT output. If so, these two words are merged into a compound, using a list-based approach, similar to Popović et al. (2006).

8 Reordering

For the system CimS-Syntax-RORI, English data parsed with EGRET was reordered using scripts written for parse trees produced by the constituent parser (Charniak and Johnson, 2005), using a model we trained on the standard Penn Treebank sections. Unfortunately, the reordering scripts could not be straightforwardly applied to EGRET parses and require more significant modifications than we first expected.

We thus decided to parse the Europarl data (v7) with (Charniak and Johnson, 2005) instead and run our reordering scripts on it (CimS-RO). For evaluation purposes, we build a baseline system *raw'* which has been trained only on Europarl. Tuning and testing setup is the same as for the systems described in Section 6 with the difference that the weights have been tuned on newstest2013. The evaluation results are shown in Table 4. Similarly to previous results reported in (Gojun and Fraser, 2012), the CimS-RO system shows an improvement of 0.5 Bleu points when compared to the *raw'* baseline .

[5]Namely: *Testflugzeugen (test airplanes), Medientribunal (media tribunal), RBS-Mitarbeiter (RBS worker), Schulmauersanierung (school wall renovation), Anti-Terror-Organisationen (anti-terror organisations), and Tabakimpfstoffe (tobacco-plant-created vaccines)* in both and in CoRI also *Hand-gepäckgebühr (hand luggage fee)* and *Haftungsstreitigkeiten (liability litigation).*

Experiment	**mert.log** news2013	**Bleu ci** news2014	**Bleu cs** news2014
raw'	16.87	16.25	15.31
CimS-RO	17.76	16.81	15.81

Table 4: Evaluation of the reordering system trained on Europarl v7.

9 Summary

We presented the CimS systems, a set of morphology-aware translation systems customised for translation from English to German. Each system operates on a different level of morphological description, be it nominal inflection, verbal inflection, compound processing or source-side reordering. Some of the systems are well-established (RI, CoRI and RO), others are still under developement (RIVe, CoRIVe and Syntax-RORI). However, all of them, with the exception of CoRIVe, lead to improved translation quality when evaluated against a contrastive baseline without linguistic processing. In an additional evaluation, we could show that the compound processing systems are able to create a considerable number of compounds unseen in the parallel training data.

In the future, we will investigate further combinations and extensions of our morphological components, including reordering, compound processing and verbal inflection. There are still many many interesting challenges to be solved in all of these areas, and this is especially true for verbal inflection.

Acknowledgments

This work was supported by Deutsche Forschungsgemeinschaft grants Models of Morphosyntax for Statistical Machine Translation (Phase 2) and Distributional Approaches to Semantic Relatedness. We would like to thank Daniel Quernheim for sharing the workload of preprocessing the data with us.

Moreover, we thank Edgar Hoch from the IMS system administration for generously providing us with disk space and all our colleagues at IMS, especially Fabienne Braune, Junfei Guo, Nina Seemann and Jason Utt for postponing their experiments to let us use most of IMS' computing facilities for a whole week. Thank you each beaucoup!

References

Fabienne Cap, Alexander Fraser, Marion Weller, and Aoife Cahill. 2014. How to Produce Unseen Teddy Bears: Improved Morphological Processing of Compounds in SMT. In *Proceedings of EACL 2014*.

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)*, Ann Arbor, Michigan.

Colin Cherry and George Foster. 2012. Batch Tuning Strategies for Statistical Machine Translation. In *Proceedings of HLT-NAACL 2012*.

Michael Collins, Philipp Koehn, and Ivona Kučerová. 2005. Clause restructuring for statistical machine translation. In *Proceedings ACL 2005*.

Chris Dyer. 2009. Using a maximum entropy model to build segmentation lattices for MT. In *Proceedings of HLT-NAACL 2009*.

Alexander Fraser, Marion Weller, Aoife Cahill, and Fabienne Cap. 2012. Modeling Inflection and Word Formation in SMT. In *Proceedings of EACL 2012*.

Alexander Fraser. 2009. Experiments in Morphosyntactic Processing for Translation to and from German. In *Proceedings of WMT 2009*.

Fabienne Fritzinger and Alexander Fraser. 2010. How to Avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In *Proceedings of WMT@ACL2010*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a Translation Rule? In *Proceedings of HLT-NAACL 2004*.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *ACL 2008: Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing*.

Adrià De Gispert and José B. Mariño. 2008. On the impact of morphology in English to Spanish statistical MT. *Speech Communication*.

Anita Gojun and Alexander Fraser. 2012. Determining the placement of German verbs in English-to-German SMT. In *Proceedings of EACL 2012*.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of WMT 2011*.

Minwoo Jeong, Kristina Toutanova, Hisami Suzuki, and Chris Quirk. 2010. A discriminative lexicon model for complex morphology. In *Proceedings of AMTA 2010*.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *Proceedings of EACL 2003*.

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 ACL 2007 (Demo Session)*.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *ICML'01*.

Klaus Macherey, Andrew M. Dai, David Talbot, Ashok C. Popat, and Franz Och. 2011. Language-independent Compound Splitting with Morphological Operations. In *Proceedings of ACL 2011*.

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 ACL 2002*.

Maja Popović, Daniel Stein, and Hermann Ney. 2006. Statistical Machine Translation of German Compound Words. In *Proceedings of FinTAL 2006*.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of COLING 2008*.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In *Proceedings of LREC 2004*.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proceedings of Coling 2004*.

Andreas Stolcke. 2002. SRILM – an Extensible Language Modelling Toolkit. In *Proceedings of ICSLN 2002*.

Sara Stymne and Nicola Cancedda. 2011. Productive Generation of Compound Words in Statistical Machine Translation. In *Proceedings of WMT@EMNLP'11*.

Sara Stymne, Maria Holmqvist, and Lars Ahrenberg. 2008. Effects of Morphological Analysis in Translation between German and English. In *Proceedings of WMT 2008*.

Sara Stymne. 2009. A Comparison of Merging Strategies for Translation of German Compounds. In *Proceedings of EACL 2009 (Student Workshop)*.

Kristina Toutanova, Hisami Suzuki, and Achim Ruopp. 2008. Applying Morphology Generation Models to Machine Translation. In *Proceedings of HLT-ACL 2008*.

Marion Weller, Alexander Fraser, and Sabine Schulte im Walde. 2013. Using Subcategorization Knowledge to Improve Case Prediction for Translation to German. In *Proceedings of ACL'13*.

Philip Williams and Philipp Koehn. 2011. Agreement constraints for statistical machine translation into German. In *Proceedings of WMT 2011*.

Philip Williams and Phillipp Koehn. 2012. GHKM-Rule Extraction and Scope-3 Parsing in Moses. In *Proceedings of WMT 2007*.