Edinburgh's Syntax-Based Systems at WMT 2014

Philip Williams[1], Rico Sennrich[1], Maria Nadejde[1],
Matthias Huck[1], Eva Hasler[1], Philipp Koehn[1,2]
[1]School of Informatics, University of Edinburgh
[2]Center for Speech and Language Processing, The Johns Hopkins University

Abstract

This paper describes the string-to-tree systems built at the University of Edinburgh for the WMT 2014 shared translation task. We developed systems for English-German, Czech-English, French-English, German-English, Hindi-English, and Russian-English. This year we improved our English-German system through target-side compound splitting, morphosyntactic constraints, and refinements to parse tree annotation; we addressed the out-of-vocabulary problem using transliteration for Hindi and Russian and using morphological reduction for Russian; we improved our German-English system through tree binarization; and we reduced system development time by filtering the tuning sets.

1 Introduction

For this year's WMT shared translation task we built syntax-based systems for six language pairs:

- English-German
- Czech-English
- French-English
- German-English
- Hindi-English
- Russian-English

As last year (Nadejde et al., 2013), our systems are based on the string-to-tree pipeline implemented in the Moses toolkit (Koehn et al., 2007).

We paid particular attention to the production of grammatical German, trying various parsers and incorporating target-side compound splitting and morphosyntactic constraints; for Hindi and Russian, we employed the new Moses transliteration model to handle out-of-vocabulary words; and for German to English, we experimented with tree binarization, obtaining good results from right binarization.

We also present our first syntax-based results for French-English, the scale of which defeated us last year. This year we were able to train a system using all available training data, a task that was made considerably easier through principled filtering of the tuning set. Although our system was not ready in time for human evaluation, we present BLEU scores in this paper.

In addition to the five single-system submissions described here, we also contributed our English-German and German-English systems for use in the collaborative EU-BRIDGE system combination effort (Freitag et al., 2014).

This paper is organised as follows. In Section 2 we describe the core setup that is common to all systems. In subsequent sections we describe language-pair specific variations and extensions. For each language pair, we present results for both the development test set (newstest2013 in most cases) and for the filtered test set (newstest2014) that was provided after the system submission deadline. We refer to these as 'devtest' and 'test', respectively.

2 System Overview

2.1 Pre-processing

The training data was normalized using the WMT `normalize-punctuation.perl` script then tokenized and truecased. Where the target language was English, we used the Moses tokenizer's `-penn` option, which uses a tokenization scheme that more closely matches that of the parser. For the English-German system we used the default Moses tokenization scheme, which is similar to that of the German parsers.

For the systems that translate into English, we used the Berkeley parser (Petrov et al., 2006; Petrov and Klein, 2007) to parse the target-side of the training corpus. As we will describe in Section 3, we tried a variety of parsers for German.

We did not perform any corpus filtering other than the standard Moses method, which removes

Proceedings of the Ninth Workshop on Statistical Machine Translation, pages 207–214,
Baltimore, Maryland USA, June 26–27, 2014.

sentence pairs with dubious length ratios and sentence pairs where parsing fails for the target-side sentence.

2.2 Translation Model

Our translation grammar is a synchronous context-free grammar (SCFG) with phrase-structure labels on the target side and the generic non-terminal label X on the source side.

The grammar was extracted from the word-aligned parallel data using the Moses implementation (Williams and Koehn, 2012) of the GHKM algorithm (Galley et al., 2004; Galley et al., 2006). For word alignment we used MGIZA++ (Gao and Vogel, 2008), a multi-threaded implementation of GIZA++ (Och and Ney, 2003).

Minimal GHKM rules were composed into larger rules subject to parameterized restrictions on size defined in terms of the resulting target tree fragment. A good choice of parameter settings depends on the annotation style of the target-side parse trees. We used the settings shown in Table 1, which were chosen empirically during the development of last years' systems:

Parameter	Value
Rule depth	5
Node count	20
Rule size	5

Table 1: Parameter settings for rule composition.

Further to the restrictions on rule composition, fully non-lexical unary rules were eliminated using the method described in Chung et al. (2011) and rules with scope greater than 3 (Hopkins and Langmead, 2010) were pruned from the translation grammar. Scope pruning makes parsing tractable without the need for grammar binarization.

2.3 Language Model

We used all available monolingual data to train 5-gram language models. Language models for each monolingual corpus were trained using the SRILM toolkit (Stolcke, 2002) with modified Kneser-Ney smoothing (Chen and Goodman, 1998) and then interpolated using weights tuned to minimize perplexity on the development set.

2.4 Feature Functions

Our feature functions are unchanged from the previous two years. They include the n-gram language model probability of the derivation's target yield, its word count, and various scores for the synchronous derivation.

Each grammar rule has a number of pre-computed scores. For a grammar rule r of the form

$$C \rightarrow \langle \alpha, \beta, \sim \rangle$$

where C is a target-side non-terminal label, α is a string of source terminals and non-terminals, β is a string of target terminals and non-terminals, and $\sim$ is a one-to-one correspondence between source and target non-terminals, we score the rule according to the following functions:

- $p\left(C, \beta \mid \alpha, \sim\right)$ and $p\left(\alpha \mid C, \beta, \sim\right)$, the direct and indirect translation probabilities.
- $p_{lex}\left(\beta \mid \alpha\right)$ and $p_{lex}\left(\alpha \mid \beta\right)$, the direct and indirect lexical weights (Koehn et al., 2003).
- $p_{pcfg}\left(\pi\right)$, the monolingual PCFG probability of the tree fragment π from which the rule was extracted.
- $\exp(-1/count(r))$, a rule rareness penalty.
- $\exp(1)$, a rule penalty. The main grammar and glue grammars have distinct penalty features.

2.5 Tuning

The feature weights were tuned using the Moses implementation of MERT (Och, 2003) for all systems except English-to-German, for which we used k-best MIRA (Cherry and Foster, 2012) due to the larger number of features.

We used tuning sentences drawn from all of the previous years' test sets (except newstest2013, which was used as the development test set). In order to speed up the tuning process, we used subsets of the full tuning sets with sentence pairs up to length 30 (Max-30) and further applied a filtering technique to reduce the tuning set size to 2,000 sentence pairs for the language pairs involving German, French and Czech[1]. We also experimented with random subsets of size 2,000.

For the filtering technique, we make the assumption that finding suitable weights for all the feature functions requires the optimizer to see a range of feature values and to see hypotheses that can partially match the reference translations in order to rank the hypotheses. For example, if a

[1] For Russian and Hindi, the development sets are smaller and no filtering was applied.

tuning example contains many out-of-vocabulary words or is difficult to translate for other reasons, this will result in low quality translation hypotheses and provide the system with little evidence for which features are useful to produce good translations. Therefore, we select high quality examples using a smooth version of sentence-BLEU computed on the 1-best output of a single decoder run on the development set. Standard sentence-BLEU tends to select short examples because they are more likely to have perfect n-gram matches with the reference translation. Very short sentence pairs are less informative for tuning but also tend to have more extreme source-target length ratios which can affect the weight of the word penalty. Thus, we penalize short examples by padding the decoder output with a fixed number of non-matching tokens[2] to the left and right before computing sentence-BLEU. This has the effect of reducing the precision of short sentences against the reference translation while affecting longer sentences proportionally less. Experiments on phrase-based systems have shown that the resulting tuning sets are of comparable diversity as randomly selected sets in terms of their feature vectors and maintain BLEU scores in comparison with tuning on the entire development set.

Table 2 shows the size of the full tuning sets and the size of the subsets with up to length 30, Table 3 shows the results of tuning with different sets. Reducing the tuning sets to Max-30 results in a speed-up in tuning time but affects the performance on some of the devtest/test sets (mostly for Czech-English). However, tuning on the full set took more than 18 days using 12 cores for German-English which is not feasible when trying out several model variations. Further filtering these subsets to a size of 2,000 sentence pairs as described above maintains the BLEU scores in most cases and even improves the scores in some cases. This indicates that the quality of the selected examples is more important than the total number of tuning examples. However, the experiments with random subsets from Max-30 show that random selection also yields results which improve over the results with Max-30 in most cases, though are not always as good as with the filtered sets.[3] The filtered tuning sets yield reasonable performance compared to the full tuning sets except for the German-English devtest set where performance drops by 0.5 BLEU[4].

Tuning set	Cs-En	En-De	De-En
Full	13,055	13,071	13,071
Max-30	10,392	9,151	10,610

Table 2: Size of full tuning sets and with sentence length up to 30.

		devtest	
Tuning set	Cs-En	En-De	De-En
Full	25.1	19.9	26.7
Max-30	24.7	19.8	26.2
Filtered	24.9	19.8	26.2
Random	24.8	19.7	26.4

		test	
Tuning set	Cs-En	En-De	De-En
Full	27.5	19.2	26.9
Max-30	27.2	19.2	27.0
Filtered	27.5	19.1	27.2
Random	27.3	19.4	27.0

Table 3: BLEU results on devtest and test sets with different tuning sets: Full, Max-30, filtered subsets of Max-30 and average of three random subsets of Max-30 (size of filtered/random subsets: 2,000).

3 English to German

We use the projective output of the dependency parser ParZu (Sennrich et al., 2013) for the syntactic annotation of our primary submission. Contrastive systems were built with other parsers: BitPar (Schmid, 2004), the German Stanford Parser (Rafferty and Manning, 2008), and the German Berkeley Parser (Petrov and Klein, 2007; Petrov and Klein, 2008).

The set of syntactic labels provided by ParZu has been refined to reduce overgeneralization phenomena. Specifically, we disambiguate the labels ROOT (used for the root of a sentence, but also commas, punctuation marks, and sentence fragments), KON and CJ (coordinations of different constituents), and GMOD (pre- or postmodifying genitive modifier).

[2] These can be arbitrary tokens that do not match any reference token.

[3] For random subsets from the full tuning set the performance was similar but resulted in standard deviations of up to 0.36 across three random sets.

[4] Note however that due to the long tuning times, we are reporting single tuning runs.

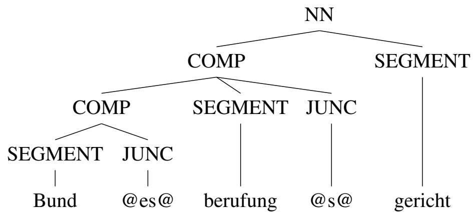

Figure 1: Syntactic representation of split compound *Bundesberufungsgericht* (Engl: *federal appeals court*).

We discriminatively learn non-terminal labels for unknown words using sparse features, rather than estimating a probability distribution of non-terminal labels from singleton statistics in the training corpus.

We perform target-side compound splitting, using a hybrid method described by Fritzinger and Fraser (2010) that combines a finite-state morphology and corpus statistics. As finite-state morphology analyzer, we use Zmorge (Sennrich and Kunz, 2014). An original contribution of our experiments is a syntactic representation of split compounds which eliminates typical problems with target-side compound splitting, namely erroneous reorderings and compound merging. We represent split compounds as a syntactic tree with the last segment as head, preceded by a modifier. A modifier consists of an optional modifier, a segment and a (possibly empty) joining element. An example is shown in Figure 1. This hierarchical representation ensures that compounds can be easily merged in post-processing (by removing the spaces and special characters around joining elements), and that no segments are placed outside of a compound in the translation.

We use unification-based constraints to model morphological agreement within German noun phrases, and between subjects and verbs (Williams and Koehn, 2011). Additionally, we add constraints that operate on the internal tree structure of the translation hypotheses, to enforce several syntactic constraints that were frequently violated in the baseline system:

- correct subcategorization of auxiliary/modal verbs in regards to the inflection of the full verb.
- passive clauses are not allowed to have accusative objects.
- relative clauses must contain a relative (or interrogative) pronoun in their first constituent.

system	BLEU	
	devtest	test
Stanford Parser	19.0	18.3
Berkeley Parser	19.3	18.6
BitPar	19.5	18.6
ParZu	19.6	19.1
+ modified label set	19.8	19.1
+ discriminative UNK weights	19.9	19.2
+ German compound splitting	20.0	19.8
+ grammatical constraints	20.2	20.1

Table 4: English to German translation results on devtest (newstest2013) and test (newstest2014) sets.

Table 4 shows BLEU scores with systems trained with different parsers, and for our extensions of the baseline system.

4 Czech to English

For Czech to English we used the core setup described in Section 2 without modification. Table 5 shows the BLEU scores.

	BLEU	
system	devtest	test
baseline	24.8	27.0

Table 5: Czech to English results on the devtest (newstest2013) and test (newstest2014) sets.

5 French to English

For French to English, alignment of the parallel corpus was performed using *fast_align* (Dyer et al., 2013) instead of MGIZA++ due to the large volume of parallel data.

Table 6 shows BLEU scores for the system and Table 7 shows the resulting grammar sizes after filtering for the evaluation sets.

	BLEU	
system	devtest	test
baseline	29.4	32.3

Table 6: French to English results on the devtest (newsdev2013) and test (newstest2014) sets.

system	devtest	test
baseline	86,341,766	88,657,327

Table 7: Grammar sizes of the French to English system after filtering for the devtest (newstest2013) and test (newstest2014) sets.

6 German to English

German compounds were split using the script provided with Moses.

For training the primary system, the target parse trees were restructured before rule extraction by *right binarization*. Since binarization strategies increase the tree depth and number of nodes by adding virtual non-terminals, we increased the extraction parameters to: *Rule Depth = 7, Node Count = 100, Rule Size = 7*. A thorough investigation of binarization methods for restructuring Penn Treebank style trees was carried out by Wang et al. (2007).

Table 8 shows BLEU scores for the baseline system and two systems employing different binarization strategies. Table 9 shows the resulting grammar sizes after filtering for the evaluation sets. Results on the development set showed no improvement when *left binarization* was used for restructuring the trees, although the grammar size increased significantly.

	BLEU	
system	devtest	test
baseline	26.2	27.2
+ right binarization (primary)	26.8	28.2
+ left binarization	26.3	-

Table 8: German to English results on the devtest (newsdev2013) and test (newstest2014) sets.

system	devtest	test
baseline	11,462,976	13,811,304
+ right binarization	24,851,982	29,133,910
+ left binarization	21,387,976	-

Table 9: Grammar sizes of the German to English systems after filtering for the devtest (newstest2013) and test (newstest2014) sets.

7 Hindi to English

English-Hindi has the least parallel training data of this year's language pairs. Out-of-vocabulary (OOV) input words are therefore a comparatively large source of translation error: in the devtest set (newsdev2014) and filtered test set (newstest2014) the average OOV rates are 1.08 and 1.16 unknown words per sentence, respectively.

Assuming a significant fraction of OOV words to be named entities and thus amenable to transliteration, we applied the post-processing transliteration method described in Durrani et al. (2014) and implemented in Moses. In brief, this is an unsupervised method that i) uses EM to induce a corpus of transliteration examples from the parallel training data; ii) learns a monotone character-level phrase-based SMT model from the transliteration corpus; and iii) substitutes transliterations for OOVs in the system output by using the monolingual language model and other features to select between transliteration candidates.[5]

Table 10 shows BLEU scores with and without transliteration on the devtest and filtered test sets. Due to a bug in the submitted system, the language model trained on the HindEnCorp corpus was used for transliteration candidate selection rather than the full interpolated language model. This was fixed subsequent to submission.

	BLEU	
system	devtest	test
baseline	12.9	14.7
+ transliteration (submission)	13.3	15.1
+ transliteration (fixed)	13.6	15.5

Table 10: Hindi to English results with and without transliteration on the devtest (newsdev2014) and test (newstest2014) sets.

Transliteration increased 1-gram precision from 48.1% to 49.4% for devtest and from 49.1% to 50.6% for test. Of the 2,913 OOV words in test, 938 (32.2%) of transliterations exactly match the reference. Manual inspection reveals that there are also many near matches. For instance, transliteration produces *Bernat Jackie* where the reference is *Jacqui Barnat*.

8 Russian to English

Compared to Hindi-English, the Russian-English language pair has over six times as much parallel data. Nonetheless, OOVs remain a problem: the average OOV rates are approximately half those

[5] This is the variant referred to as Method 2 in Durrani et al. (2014).

of Hindi-English, at 0.47 and 0.51 unknown words per sentence for the devtest (newstest2013) and filtered test (newstest2014) sets, respectively. We address this in part using the same transliteration method as for Hindi-English.

Data sparsity issues for this language pair are exacerbated by the rich inflectional morphology of Russian. Many Russian word forms express grammatical distinctions that are either absent from English translations (like grammatical gender) or are expressed by different means (like grammatical function being expressed through syntactic configuration rather than case). We adopt the widely-used approach of simplifying morphologically-complex source forms to remove distinctions that we believe to be redundant. Our method is similar to that of Weller et al. (2013) except that ours is much more conservative (in their experiments, Weller et al. (2013) found morphological reduction to harm translation indicating that useful information was likely to have been discarded).

We used TreeTagger (Schmid, 1994) to obtain a lemma-tag pair for each Russian word. The tag specifies the word class and various morphosyntactic feature values. For example, the adjective республиканская ('republican') gets the lemma-tag pair республиканский + `Afpfsnf`, where the code A indicates the word class and the remaining codes indicate values for the type, degree, gender, number, case, and definiteness features.

Like Weller et al. (2013), we selectively replaced surface forms with their lemmas and reduced tags, reducing tags through feature deletion. We restricted morphological reduction to adjectives and verbs, leaving all other word forms unchanged. Table 11 shows the features that were deleted. We focused on contextual inflection, making the assumption that inflectional distinctions required by agreement alone were the least likely to be useful for translation (since the same information was marked elsewhere in the sentence) and also the most likely to be the source of 'spurious' variation.

Table 12 shows the BLEU scores for Russian-English with transliteration and morphological reduction. The effect of transliteration was smaller than for Hindi-English, as might be expected from the lower baseline OOV rate. 1-gram precision increased from 57.1% to 57.6% for devtest and from 62.9% to 63.6% for test. Morphological reduction decreased the initial OOV rates by 3.5% and 4.1% on the devtest and filtered test sets. After both morphological and transliteration the 1-gram precisions for devtest and test were 57.7% and 63.8%.

Adjective		Verb	
Type	✗	Type	✗
Degree	✓	VForm	✓
Gender	✗	Tense	✓
Number	✗	Person	✓
Case	✗	Number	✓
Definiteness	✗	Gender	✗
		Voice	✓
		Definiteness	✗
		Aspect	✓
		Case	✓

Table 11: Feature values that are retained (✓) or deleted (✗) during morphological reduction of Russian.

	BLEU	
system	devtest	test
baseline	23.3	29.7
+ transliteration	23.7	30.3
+ morphological reduction	23.8	30.3

Table 12: Russian to English results on the devtest (newstest2013) and test (newstest2014) sets.

9 Conclusion

We have described Edinburgh's syntax-based systems in the WMT 2014 shared translation task. Building upon the already-strong string-to-tree systems developed for previous years' shared translation tasks, we have achieved substantial improvements over our baseline setup: we improved translation into German through target-side compound splitting, morphosyntactic constraints, and refinements to parse tree annotation; we have addressed unknown words using transliteration (for Hindi and Russian) and morphological reduction (for Russian); and we have improved our German-English system through tree binarization.

Acknowledgements

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n^{o} 287658 (EU-BRIDGE).

Rico Sennrich has received funding from the Swiss National Science Foundation under grant P2ZHP1_148717.

References

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical report, Harvard University.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 427–436, Montréal, Canada, June. Association for Computational Linguistics.

Tagyoung Chung, Licheng Fang, and Daniel Gildea. 2011. Issues concerning decoding with synchronous context-free grammar. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 413–417, Portland, Oregon, USA, June.

Nadir Durrani, Hassan Sajjad, Hieu Hoang, and Philipp Koehn. 2014. Integrating an Unsupervised Transliteration Model into Statistical Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the ACL (EACL 2014)*, Gothenburg, Sweden, April. To appear.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of ibm model 2. In *In Proc. NAACL/HLT 2013*, pages 644–648.

Markus Freitag, Stephan Peitz, Joern Wuebker, Hermann Ney, Matthias Huck, Rico Sennrich, Nadir Durrani, Maria Nadejde, Philip Williams, Philipp Koehn, Teresa Herrmann, Eunah Cho, and Alex Waibel. 2014. EU-BRIDGE MT: Combined Machine Translation. In *Proceedings of the ACL 2014 Ninth Workshop on Statistical Machine Translation*, Baltimore, MD, USA, June.

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 Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, WMT '10, pages 224–234, Uppsala, Sweden.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a Translation Rule? In *HLT-NAACL '04*.

Michel Galley, Jonathan Graehl, Kevin Knight, Daniel Marcu, Steve DeNeefe, Wei Wang, and Ignacio Thayer. 2006. Scalable inference and training of context-rich syntactic translation models. In *ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 961–968, Morristown, NJ, USA.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57, Stroudsburg, PA, USA.

Mark Hopkins and Greg Langmead. 2010. SCFG decoding without binarization. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 646–655, Cambridge, MA, October.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 48–54, Morristown, NJ, USA.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Morristown, NJ, USA. Association for Computational Linguistics.

Maria Nadejde, Philip Williams, and Philipp Koehn. 2013. Edinburgh's Syntax-Based Machine Translation Systems. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 170–176, Sofia, Bulgaria, August.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Morristown, NJ, USA.

Slav Petrov and Dan Klein. 2007. Improved Inference for Unlexicalized Parsing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 404–411, Rochester, New York, April.

Slav Petrov and Dan Klein. 2008. Parsing German with Latent Variable Grammars. In *Proceedings of the Workshop on Parsing German at ACL '08*, pages 33–39, Columbus, OH, USA, June.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440.

Anna N. Rafferty and Christopher D. Manning. 2008. Parsing Three German Treebanks: Lexicalized and Unlexicalized Baselines. In *Proceedings of the*

Workshop on Parsing German at ACL '08, pages 40–46, Columbus, OH, USA, June.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Helmut Schmid. 2004. Efficient Parsing of Highly Ambiguous Context-Free Grammars with Bit Vectors. In *Proc. of the Int. Conf. on Computational Linguistics (COLING)*, Geneva, Switzerland, August.

Rico Sennrich and Beat Kunz. 2014. Zmorge: A German Morphological Lexicon Extracted from Wiktionary. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, May.

Rico Sennrich, Martin Volk, and Gerold Schneider. 2013. Exploiting Synergies Between Open Resources for German Dependency Parsing, POS-tagging, and Morphological Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2013*, pages 601–609, Hissar, Bulgaria.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Intl. Conf. Spoken Language Processing, Denver, Colorado, September 2002.*

Wei Wang, Kevin Knight, Daniel Marcu, and Marina Rey. 2007. Binarizing Syntax Trees to Improve Syntax-Based Machine Translation Accuracy. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 746–754.

Marion Weller, Max Kisselew, Svetlana Smekalova, Alexander Fraser, Helmut Schmid, Nadir Durrani, Hassan Sajjad, and Richárd Farkas. 2013. Munich-Edinburgh-Stuttgart submissions at WMT13: Morphological and syntactic processing for SMT. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 232–239, Sofia, Bulgaria, August.

Philip Williams and Philipp Koehn. 2011. Agreement Constraints for Statistical Machine Translation into German. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 217–226, Edinburgh, Scotland, July.

Philip Williams and Philipp Koehn. 2012. GHKM Rule Extraction and Scope-3 Parsing in Moses. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 388–394, Montréal, Canada, June.