A Joint Dependency Model of Morphological and Syntactic Structure for Statistical Machine Translation

Rico Sennrich and **Barry Haddow**
School of Informatics, University of Edinburgh
rico.sennrich@ed.ac.uk, bhaddow@inf.ed.ac.uk

Abstract

When translating between two languages that differ in their degree of morphological synthesis, syntactic structures in one language may be realized as morphological structures in the other, and SMT models need a mechanism to learn such translations. Prior work has used morpheme splitting with flat representations that do not encode the hierarchical structure between morphemes, but this structure is relevant for learning morphosyntactic constraints and selectional preferences. We propose to model syntactic and morphological structure jointly in a dependency translation model, allowing the system to generalize to the level of morphemes. We present a dependency representation of German compounds and particle verbs that results in improvements in translation quality of 1.4–1.8 BLEU in the WMT English–German translation task.

1 Introduction

When translating between two languages that differ in their degree of morphological synthesis, syntactic structures in one language may be realized as morphological structures in the other. Machine Translation models that treat words as atomic units have poor learning capabilities for such translation units, and morphological segmentations are commonly used (Koehn and Knight, 2003). Like words in a sentence, the morphemes of a word have a hierarchical structure that is relevant in translation. For instance, compounds in Germanic languages are head-final, and the head is the segment that determines agreement within the noun phrase, and is relevant for selectional preferences of verbs.

1. sie erheben eine Hand|gepäck|gebühr.
   they charge a carry-on bag fee.

function/postion	English/German example
finite (main)	he **walks away** quickly er **geht** schnell **weg**
finite (sub.)	[...] because he **walks away** quickly [...] weil er schnell **weggeht**
bare infinitive	he can **walk away** quickly er kann schnell **weggehen**
to/zu-infinitive	he promises **to walk away** quickly er verspricht, schnell **wegzugehen**

Table 1: Surface realizations of particle verb *weggehen* 'walk away'.

In example 1, agreement in case, number and gender is enforced between *eine* 'a' and *Gebühr* 'fee', and selectional preference between *erheben* 'charge' and *Gebühr* 'fee'. A flat representation, as is common in phrase-based SMT, does not encode these relationships, but a dependency representation does so through dependency links.

In this paper, we investigate a dependency representation of morphologically segmented words for SMT. Our representation encodes syntactic and morphological structure jointly, allowing a single model to learn the translation of both. Specifically, we work with a string-to-tree model with GHKM-style rules (Galley et al., 2006), and a relational dependency language model (Sennrich, 2015). We focus on the representation of German syntax and morphology in an English-to-German system, and two morphologically complex word classes in German that are challenging for translation, compounds and particle verbs.

German makes heavy use of compounding, and compounds such as *Abwasserbehandlungsanlage* 'waste water treatment plant' are translated into complex noun phrases in other languages, such as French *station d'épuration des eaux résiduaires*.

German particle verbs are difficult to model because their surface realization differs depending on the finiteness of the verb and the type of clause. Verb particles are separated from the finite verb in

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2081–2087,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

main clauses, but prefixed to the verb in subordinated clauses, or when the verb is non-finite. The infinitive marker *zu* 'to', which is normally a premodifying particle, appears as an infix in particle verbs. Table 1 shows an illustrating example.

2 A Dependency Representation of Compounds and Particle Verbs

The main focus of research on compound splitting has been on the splitting algorithm (Popovic et al., 2006; Nießen and Ney, 2000; Weller et al., 2014; Macherey et al., 2011). Our focus is not the splitting algorithm, but the representation of compounds. For splitting, we use an approach similar to (Fritzinger and Fraser, 2010), with segmentation candidates identified by a finite-state morphology (Schmid et al., 2004; Sennrich and Kunz, 2014), and statistical evidence from the training corpus to select a split (Koehn and Knight, 2003).

German compounds are head-final, and premodifiers can be added recursively. Compounds are structurally ambiguous if there is more than one modifier. Consider the distinction between *(Stadtteil)projekt* (literally: '(city part) project)') and *Stadt(teilprojekt)* 'city sub-project'. We opt for a left-branching representation by default.[1] We also split linking elements, and represent them as a postmodifier of each non-final segment, including the empty string ("ϵ"). We use the same representation for noun compounds and adjective compounds.

An example of the original[2] and the proposed compound representation is shown in Figure 1. Importantly, the head of the compound is also the parent of the determiners and attributes in the noun phrase, which makes a bigram dependency language model sufficient to enforce agreement. Since we model morphosyntactic agreement within the main translation step, and not in a separate step as in (Fraser et al., 2012), we deem it useful that inflection is marked at the head of the compound. Consequently, we do not split off inflectional or derivational morphemes.

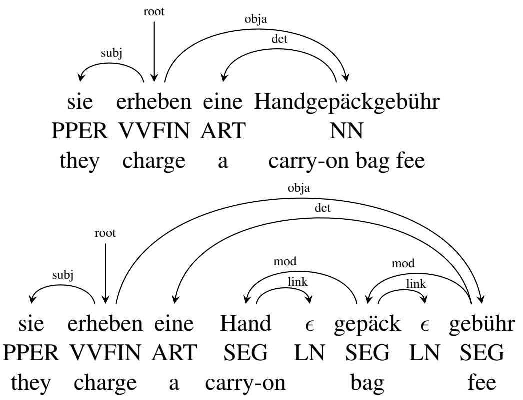

root
subj
obja
det
sie erheben eine Handgepäckgebühr
PPER VVFIN ART NN
they charge a carry-on bag fee
root
subj
obja
det
mod
link
mod
link
sie erheben eine Hand ϵ gepäck ϵ gebühr
PPER VVFIN ART SEG LN SEG LN SEG
they charge a carry-on bag fee

Figure 1: Original and proposed representation of German compound.

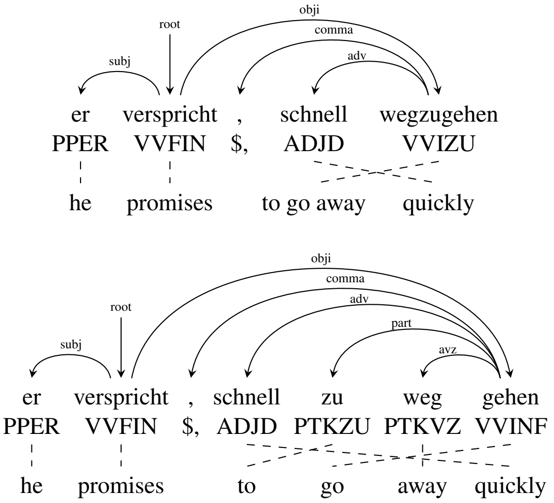

root
subj
obji
comma
adv
er verspricht , schnell wegzugehen
PPER VVFIN $, ADJD VVIZU
he promises to go away quickly
root
subj
obji
comma
adv
part
avz
er verspricht , schnell zu weg gehen
PPER VVFIN $, ADJD PTKZU PTKVZ VVINF
he promises to go away quickly

Figure 2: Original and proposed representation of German particle verb with infixed *zu*-marker.

For German particle verbs, we define a common representation that abstracts away from the various surface realizations (see Table 1). Separated verb particles are reordered to be the closest premodifier of the verb. Prefixed particles and the *zu*-infix are identified by the finite-state-morphology, and split from the verb so that the particle is the closest, the *zu* marker the next-closest premodifier of the verb, as shown in Figure 2. Agreement, selectional preferences, and other phenomena involve the verb and its dependents, and the proposed representation retains these dependency links, but reduces data sparsity from affixation and avoids discontinuity of the verb and its particle.

3 Tree Binarization

We follow Williams et al. (2014) and map dependency trees into a constituency representation, which allows for the extraction of GHKM-style translation rules (Galley et al., 2006). This conversion is lossless, and we can still apply a de-

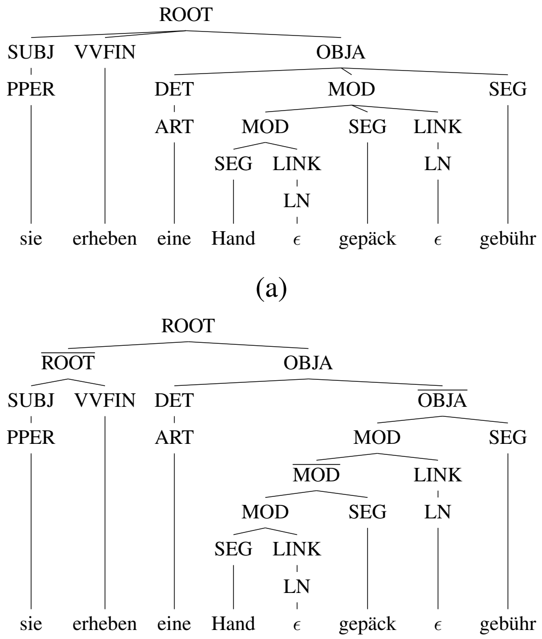
[1] We follow prior work in leaving frequent words or subwords unsplit, which has a disambiguating effect. With more aggressive splitting, frequency information could be used for the structural disambiguation of internal structure.

[2] The original dependency trees follow the annotation guidelines by Foth (2005).

pendency language model (RDLM). Figure 3 (a) shows the constituency representation of the example in Figure 1.

Our model should not only be able to produce new words productively, but also to memorize words it has observed during training. Looking at the compound *Handgepäckgebühr* in Figure 3 (a), we can see that it does not form a constituent, and cannot be extracted with GHKM extraction heuristics. To address this, we binarize the trees in our training data (Wang et al., 2007).

A complicating factor is that the binarization should not impair the RDLM. During decoding, we map the internal tree structure of each hypothesis back to the unbinarized form, which is then scored by the RDLM. Virtual nodes introduced by the binarization must also be scorable by RDLM if they form the root of a translation hypothesis. A simple right or left binarization would produce virtual nodes without head and without meaningful dependency representation. We ensure that each virtual node dominates the head of the full constituent through a mixed binarization.[3] Specifically, we perform right binarization of the head and all pre-modifiers, then left binarization of all post-modifiers. This head-binarized representation is illustrated in Figure 3 (b).[4]



(a)

(b)

Figure 3: Unbinarized (a) and head-binarized (b) constituency representation of Figure 1.

Head binarization ensures that even hypotheses whose root is a virtual node can be scored by the RDLM. This score is only relevant for pruning, and discarded when the full constituent is scored. Still, these hypotheses require special treatment in the RDLM to mitigate search errors. The virtual node labels (such as $\overline{\text{OBJA}}$) are unknown symbols to the RDLM, and we simply replace them with the original label (OBJA). The RDLM uses sibling context, and this is normally padded with special start and stop symbols, analogous to BOS/EOS symbols in n-gram models. These start and stop symbols let the RDLM compute the probability that a node is the first or last child of its ancestor node. However, computing these probabilities for virtual nodes would unfairly bias the search, since the first/last child of a virtual node is not necessarily the first/last child of the full constituent. We adapt the representation of virtual nodes in RDLM to take this into account. We distinguish between virtual nodes based on whether their span is a string prefix, suffix, or infix of the full constituent. For prefixes and infixes, we do not add a stop symbol at the end, and use null symbols, which denote unavailable context, for padding to the right. For suffixes and infixes, we do the same at the start.

4 Post-Processing

For SMT, all German training and development data is converted into the representation described in sections 2–3. To restore the original representation, we start from the tree output of the string-to-tree decoder. Merging compounds is trivial: all segments and linking elements can be identified by the tree structure, and are concatenated.

For verbs that dominate a verb particle, the original order is restored through three rules:

1. non-finite verbs are concatenated with the particle, and *zu*-markers are infixed.

2. finite verbs that head a subordinated clause (identified by its dependency label) are concatenated with the particle.

3. finite verbs that head a main clause have the

[3] In other words, every node is a fixed well-formed dependency structure (Shen et al., 2010) with our binarization.

[4] Note that our definition of head binarization is different from that of Wang et al. (2007), who left-binarize a node if the head is the first child, and right-binarize otherwise. Our algorithm also covers cases where the head has both pre- and post-modifiers, as *erheben* and *gepäck* do in Figure 3.

particle moved to the right clause bracket.[5]

Previous work on particle verb translation into German proposed to predict the position of particles with an n-gram language model (Nießen and Ney, 2001). Our rules have the advantage that they are informed by the syntax of the sentence and consider the finiteness of the verb.

Our rules only produce projective trees. Verb particles may also appear in positions that violate projectivity, and we leave it to future research to determine if our limitation to projective trees affects translation quality, and how to produce non-projective trees.

5 SMT experiments

5.1 Data and Models

We train English–German string-to-tree SMT systems on the training data of the shared translation task of the Workshop on Statistical Machine Translation (WMT) 2015. The data set consists of 4.2 million sentence pairs of parallel data, and 160 million sentences of monolingual German data.

We base our systems on that of Williams et al. (2014). It is a string-to-tree GHKM translation system implemented in Moses (Koehn et al., 2007), and using the dependency annotation by ParZu (Sennrich et al., 2013). Additionally, our baseline system contains a dependency language model (RDLM) (Sennrich, 2015), trained on the target-side of the parallel training data.

We report case-sensitive BLEU scores on the newstest2014/5 test sets from WMT, averaged over 3 optimization runs of k-batch MIRA (Cherry and Foster, 2012) on a subset of newstest2008-12.[6]

We split all particle verbs and hyphenated compounds, but other compounds are only split if they are rare (frequency in parallel text < 5).

For comparison with the state-of-the-art, we train a *full system* on our restructured representation, which incorporates all models and settings of our WMT 2015 submission system (Williams et al., 2015).[7] Note that our WMT 2015 submission uses the dependency representation of compounds and tree binarization introduced in this paper; we achieve additional gains over the submission system through particle verb restructuring.

system	newstest2014	newstest2015
baseline	20.7	22.0
+split compounds	21.3	22.4
+particle verbs	21.4	22.8
head binarization	20.9	22.7
+split compounds	22.0	23.4
+particle verbs	22.1	23.8
full system	22.6	24.4

Table 2: English–German translation results (BLEU). Average of three optimization runs.

system	compound	particle verb		
		sep.	pref.	*zu*-infix
reference	2841	553	1195	176
baseline	845	96	847	71
+head binarization	798	157	858	106
+split compounds	1850	160	877	94
+particle verbs	1992	333	953	169

Table 3: Number of compounds [that would be split by compound splitter] and particle verbs (separated, prefixed and with *zu*-infix) in newstest2014/5. Average of three optimization runs.

5.2 SMT Results

Table 2 shows translation quality (BLEU) with different representations of German compounds and particle verbs. Head binarization not only yields improvements over the baseline, but also allows for larger gains from morphological segmentation. We attribute this to the fact that full compounds, and prefixed particle verbs, are not always a constituent in the segmented representation, and that binarization compensates this theoretical drawback.

With head binarization, we find substantial improvements from compound splitting of 0.7–1.1 BLEU. On newstest2014, the improvement is almost twice of that reported in related work (Williams et al., 2014), which also uses a hierarchical representation of compounds, albeit one that does not allow for dependency modelling. Examples of correct, unseen compounds generated include *Staubsauger|roboter* 'vacuum cleaner robot', *Gravitation|s|wellen* 'gravitational waves', and *NPD|-|verbot|s|verfahren* 'NPD banning process'.[8]

[5] We use the last position in the clause as default location, but put the particle before any subordinated and coordinated clauses, which occur in the *Nachfeld* (the 'final field' in topological field theory).

[6] We use *mteval-v13a.pl* for comparability to official WMT results; all significance values reported are obtained with MultEval (Clark et al., 2011).

[7] In contrast to our other systems in this paper, RDLM is trained on all monolingual data for the full system, and two models are added: a 5-gram Neural Network language model (Vaswani et al., 2013), and soft source-syntactic constraints (Huck et al., 2014).

[8] Note that *Staubsauger*, despite being a compound, is not

Particle verb restructuring yields additional gains of 0.1–0.4 BLEU. One reason for the smaller effect of particle verb restructuring is that the difficult cases – separated particle verbs and those with infixation – are rarer than compounds, with 2841 rare compounds [that would be split by our compound splitter] in the reference texts, in contrast to 553 separated particle verbs, and 176 particle verbs with infixation, as Table 3 illustrates. If we only evaluate the sentences containing a particle verb with *zu*-infix in the reference, 165 in total for newstest2014/5, we observe an improvement of 0.8 BLEU on this subset (22.1→22.9), significant with $p < 0.05$.

The positive effect of restructuring is also apparent in frequency statistics. Table 3 shows that the baseline system severely undergenerates compounds and separated/infixed particle verbs. Binarization, compound splitting, and particle verb restructuring all contribute to bringing the distribution of compounds and particle verbs closer to the reference.

In total, the restructured representation yields improvements of 1.4–1.8 BLEU over our baseline. The *full system* is competitive with official submissions to the WMT 2015 shared translation tasks. It outperforms our submission (Williams et al., 2015) by 0.4 BLEU, and outperforms other phrase-based and syntax-based submissions by 0.8 BLEU or more. The best reported result according to BLEU is an ensemble of Neural MT systems (Jean et al., 2015), which achieves 24.9 BLEU. In the human evaluation, both our submission and the Neural MT system were ranked 1–2 (out of 16), with no significant difference between them.

5.3 Synthetic LM Experiment

We perform a synthetic experiment to test our claim that a dependency representation allows for the modelling of agreement between morphemes. For 200 rare compounds [that would be split by our compound splitter] in the newstest2014/5 references, we artificially introduce agreement errors by changing the gender of the determiner. For instance, we create the erroneous sentence *sie erheben ein Handgepäckgebühr* as a complement to Example 1. We measure the ability of language models to prefer (give a higher probability to) the original reference sentence over the erroneous one. In the original representation, both a Kneser-Ney 5-gram LM and RDLM perform poorly due to data sparseness, with 70% and 57.5% accuracy, respectively. In the split representation, the RDLM reliably prefers the correct agreement (96.5% accuracy), whilst the performance of the 5-gram model even deteriorates (to 60% accuracy). This is because the gender of the first segment(s) is irrelevant, or even misleading, for agreement. For instance, *Handgepäck* is neuter, which could lead a morpheme-level n-gram model to prefer the determiner *ein*, but *Handgepäckgebühr* is feminine and requires *eine*.

6 Conclusion

Our main contribution is that we exploit the hierarchical structure of morphemes to model them jointly with syntax in a dependency-based string-to-tree SMT model. We describe the dependency annotation of two morphologically complex word classes in German, compounds and particle verbs, and show that our tree representation yields improvements in translation quality of 1.4–1.8 BLEU in the WMT English–German translation task.[9]

The principle of jointly representing syntactic and morphological structure in dependency trees can be applied to other language pairs, and we expect this to be helpful for languages with a high degree of morphological synthesis. However, the annotation needs to be adapted to the respective languages. For example, French compounds such as *arc-en-ciel* 'rainbow' are head-initial, in contrast to head-final Germanic compounds.

Acknowledgments

This project received funding from the European Union's Horizon 2020 research and innovation programme under grant agreements 645452 (QT21), 644402 (HimL), 644333 (TraMOOC), and from the Swiss National Science Foundation under grant P2ZHP1_148717.

References

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*, NAACL HLT '12, pages 427–436, Montreal, Canada. Association for Computational Linguistics.

segmented due to its frequency.

[9] We released source code and configuration files at `https://github.com/rsennrich/wmt2014-scripts`.

Jonathan H. Clark, Chris Dyer, Alon Lavie, and Noah A. Smith. 2011. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 176–181, Portland, Oregon. Association for Computational Linguistics.

Killian A. Foth. 2005. *Eine umfassende Constraint-Dependenz-Grammatik des Deutschen*. University of Hamburg, Hamburg.

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*, pages 664–674, Avignon, France. Association for Computational Linguistics.

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. Association for Computational Linguistics.

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, Sydney, Australia. Association for Computational Linguistics.

Matthias Huck, Hieu Hoang, and Philipp Koehn. 2014. Preference Grammars and Soft Syntactic Constraints for GHKM Syntax-based Statistical Machine Translation. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 148–156, Doha, Qatar. Association for Computational Linguistics.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. 2015. Montreal Neural Machine Translation Systems for WMT'15 . In *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

Philipp Koehn and Kevin Knight. 2003. Empirical Methods for Compound Splitting. In *EACL '03: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, pages 187–193, Budapest, Hungary. Association for Computational Linguistics.

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 ACL-2007 Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Klaus Macherey, Andrew Dai, David Talbot, Ashok 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: Human Language Technologies*, pages 1395–1404, Portland, Oregon, USA. Association for Computational Linguistics.

Sonja Nießen and Hermann Ney. 2000. Improving SMT quality with morpho-syntactic analysis. In *18th Int. Conf. on Computational Linguistics*, pages 1081–1085.

Sonja Nießen and Hermann Ney. 2001. Morpho-syntactic analysis for Reordering in Statistical Machine Translation. In *Machine Translation Summit*, pages 247–252, Santiago de Compostela, Spain.

Maja Popovic, Daniel Stein, and Hermann Ney. 2006. Statistical Machine Translation of German Compound Words. In *Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006*, pages 616–624, Turku, Finland.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004)*, pages 1263–1266.

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.

Rico Sennrich. 2015. Modelling and Optimizing on Syntactic N-Grams for Statistical Machine Translation. *Transactions of the Association for Computational Linguistics*, 3:169–182.

Libin Shen, Jinxi Xu, and Ralph Weischedel. 2010. String-to-dependency Statistical Machine Translation. *Comput. Linguist.*, 36(4):649–671.

Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. 2013. Decoding with Large-Scale Neural Language Models Improves Translation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*, pages 1387–1392, Seattle, Washington, USA.

Wei Wang, Kevin Knight, and Daniel Marcu. 2007. Binarizing Syntax Trees to Improve Syntax-Based Machine Translation Accuracy. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Marion Weller, Fabienne Cap, Stefan Müller, Sabine Schulte im Walde, and Alexander Fraser. 2014. Distinguishing Degrees of Compositionality in Compound Splitting for Statistical Machine Translation. In *Proceedings of the First Workshop on Computational Approaches to Compound Analysis (ComAComA 2014)*, pages 81–90, Dublin, Ireland.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, Eva Hasler, and Philipp Koehn. 2014. Edinburgh's Syntax-Based Systems at WMT 2014. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 207–214, Baltimore, Maryland, USA. Association for Computational Linguistics.

Philip Williams, Rico Sennrich, Maria Nadejde, Matthias Huck, and Philipp Koehn. 2015. Edinburgh's Syntax-Based Systems at WMT 2015. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal. Association for Computational Linguistics.