Hybrid Morphological Segmentation for Phrase-Based Machine Translation

Stig-Arne Grönroos
Department of Signal Processing and Acoustics
Aalto University, Finland
stig-arne.gronroos@aalto.fi

Sami Virpioja
Department of Computer Science
Aalto University, Finland
sami.virpioja@aalto.fi

Mikko Kurimo
Department of Signal Processing and Acoustics
Aalto University, Finland
mikko.kurimo@aalto.fi

Abstract

This article describes the Aalto University entry to the English-to-Finnish news translation shared task in WMT 2016. Our segmentation method combines the strengths of rule-based and unsupervised morphology. We also attempt to correct errors in the boundary markings by post-processing with a neural morph boundary predictor.

1 Introduction

Using words as translation tokens is problematic for synthetic languages with rich inflection, derivation or compounding. Such languages have very large vocabularies, leading to sparse statistics and many out-of-vocabulary words. Differences in morphological complexity between source and target languages also complicate alignment.

A common method for alleviating these problems is to segment the morphologically richer side as a pre-processing step. Over-segmentation is detrimental, however, as longer windows of history need to be used, and useful phrases become more difficult to extract. It is therefore important to find a balance in the amount of segmentation.

We consider the case that there are linguistic gold standard segmentations available for the morphologically complex target language. Even if there is no rule-based morphological analyzer for the language, a limited set of gold standard segmentations can be used for training a reasonably accurate statistical segmentation model in a supervised or semi-supervised manner (Ruokolainen et al., 2014; Cotterell et al., 2015).

While using a linguistically accurate morphological segmentation in a phrase-based SMT system may sound like a good idea, there is evidence that shows otherwise. In general, over-segmentation seems to be a larger problem for NLP applications than under-segmentation (Virpioja et al., 2011). In the case of SMT, linguistic morphs may provide too high granularity compared to the second language, and deteriorate alignment (Habash and Sadat, 2006; Chung and Gildea, 2009; Clifton and Sarkar, 2011). Moreover, longer sequences of units are needed in the language model and the translation phrases to cover the same span of text.

An unsupervised morphological segmentation may alleviate these problems. A method based on optimizing the training data likelihood, such as Morfessor (Creutz and Lagus, 2002; Creutz and Lagus, 2007; Virpioja et al., 2013), ensures that common phenomena are modeled more accurately, for example by using full forms for highly-frequent words even if they consist of multiple morphemes. Data-driven methods also allow tuning the segmentation granularity, for example based on symmetry between the languages in a parallel corpus (Grönroos et al., 2015).

To combine the advantages of linguistic segmentation and data-driven segmentation, we propose a hybrid approach for morphological segmentation. We optimize the segmentation in a data-driven manner, aiming for a similar granularity as the second language of the language pair, but restricting the possible set of segmentation boundaries to those between linguistic morphs. That is, the segmentation method may decide to join any of the linguistic morphs, but it cannot add new segmentation boundaries to known linguistic morphs.

We show that it is possible to improve on the linguistically accurate segmentation by reducing the amount of segmentation in an unsupervised manner.

1.1 Related work

Rule-based and statistical segmentation for SMT have been extensively studied in isolation (Virpi-

Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers, pages 289–295,
Berlin, Germany, August 11-12, 2016.

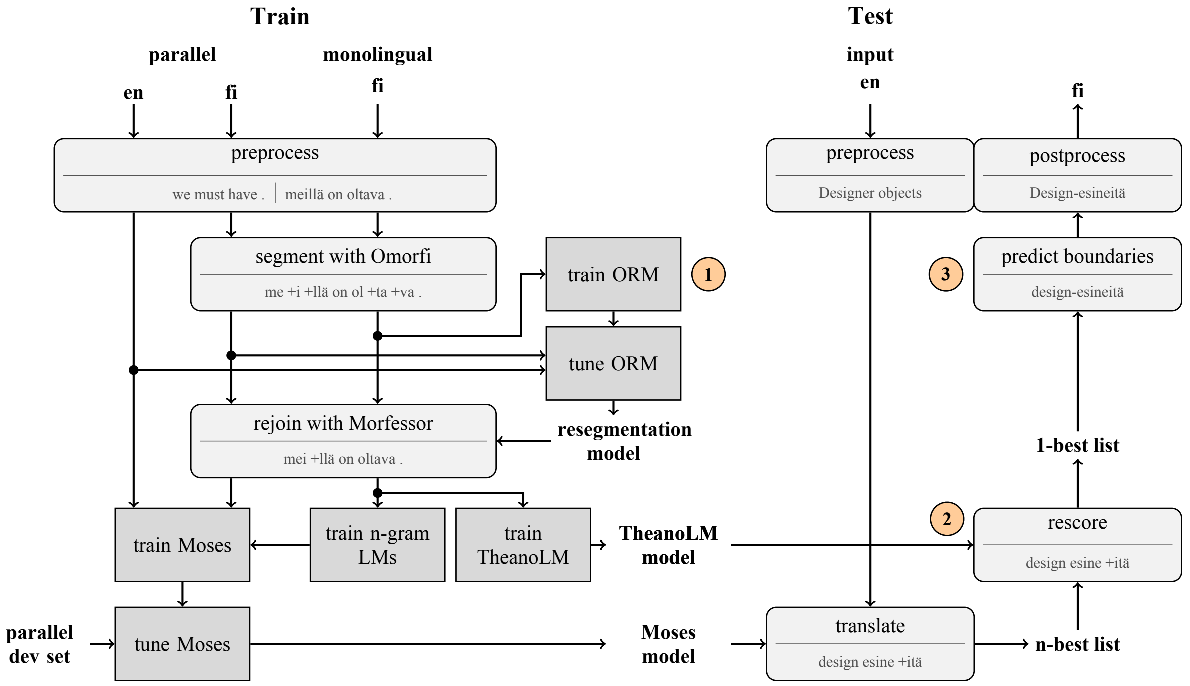

Figure 1: A pipeline overview of training of the system and using it for translation. Main contributions are hilighted with numbers 1-3. ORM is short for Omorfi-restricted Morfessor.

oja et al., 2007; Fishel and Kirik, 2010; Luong et al., 2010), and also the use of system combination to combine their strengths has been examined (De Gispert et al., 2009; Rubino et al., 2015; Pirinen et al., 2016).

Prediction of morph boundary types has been used in conjunction with compound splitting. Stymne and Cancedda (2011) apply rule-based compound splitting in the pre-processing stage, and a conditional random field with rich linguistic features for generating novel compounds in post-processing. Coalescence of compound parts in the translation output is promoted using POS-tag features. Cap et al. (2014) extend the post-predictor to also inflect the compound modifiers e.g. to add a linking morpheme.

Stymne et al. (2013) investigate several methods for splitting and merging compounds when translating into Germanic languages, and provide an extensive reading list on the topic.

2 System overview

An overview of the system is shown in Figure 1. The three main contributions of this work are indicated by numbered circles:

1. Combining rule-based morphological segmentation (Omorfi) to data-driven morphological segmentation (Morfessor).

2. Rescoring n-best lists with TheanoLM (Enarvi and Kurimo, 2016).

3. Correcting boundary markings with post-processing predictor.

Our system extends the phrase-based SMT system Moses (Koehn et al., 2007) to perform segmented translation, by adding pre-processing and post-processing steps, with no changes to the decoder.

The standard pre-processing steps not specified in Figure 1 consist of normalization of punctuation, tokenization, and statistical truecasing. All of these were performed with the tools included in Moses. The pre-processing steps are followed by morphological segmentation.

In addition, the parallel data was cleaned and duplicate sentences were removed. Cleaning was performed after morphological segmentation, as the segmentation can increase the length in tokens of a sentence.

The post-processing steps include rescoring of the n-best list, boundary prediction and desegmentation. These are followed by the standard post-processing steps, reversing the pre-processing steps: detruecasing and detokenization.

System	Tokens	Segmentation		
Words	3	hyötyajoneuvojen	tekniset	tienvarsitarkastukset
		[commercial vehicles']	*[technical]*	*[roadside inspections]*
Omorfi	11	hyöty@ ajo@ neuvo +j +en	teknise +t	tien@ varsi@ tarkastukse +t
		[utility] [drive] [counsel] [+Pl] [+Gen]	*[technical] [+Pl]*	*[road] [side] [inspection] [+Pl]*
ORM	5	hyötyajoneuvo +jen	tekniset	tienvarsi@ tarkastukset
		[commercial vehicle] [+Pl +Gen]	*[technical]*	*[roadside] [inspections]*
Source	6	technical roadside inspection of commercial vehicles		

Table 1: Worked example of two-stage morphological segmentation, beginning with rule-based Omorfi segmentation and followed by Omorfi-restricted Morfessor (ORM). The glosses below the segmentations show approximate meaning of the segments (Pl = plural suffix, Gen = genitive suffix).

2.1 Morphological segmentation

An example of the morphological segmentation is shown in Table 1.

2.1.1 Omorfi segmentation

We begin the morphological segmentation by applying the segmentation tool from Omorfi (Pirinen, 2015). Hyphens removed by Omorfi are reintroduced.

Omorfi outputs 5 types of intra-word boundaries, which we mark in different ways. Compound modifiers, identified by the `WB` or `wB` boundary type, are marked with a reserved symbol '@' at the right edge of the morph. Suffixes, identified by a leading morph boundary `MB` or derivation boundary `DB`, are marked with a '+' at the left edge. Boundaries of the type `STUB` (other stemmer-type boundary) are removed. This marking scheme leaves the compound head, or last stem of the word, unmarked. E.g. "yli`{WB}`voimai`{STUB}`s`{MB}`i`{MB}`a" is marked as "yli@ voimais +i +a".

Words not identified by Omorfi are collected in a separate vocabulary, and treated as unsegmentable.

2.1.2 Restricted Morfessor Baseline

In order to force the Morfessor method to follow the linguistic morphs produced by Omorfi, we added some new features to the Morfessor Baseline implementation by Virpioja et al. (2013). The new extension, Restricted Morfessor Baseline, is able to remove any of the given intra-word boundaries, but cannot introduce any new ones.

The standard training algorithm of Morfessor iterates over the word forms, testing whether to split the corresponding string to two parts or leave it as it is. If the string is split, the testing descends recursively to the substrings. The segmentation decisions are stored in a binary tree structure, where each node corresponds to a string. The root nodes are full word forms and leaf nodes are morphs.

The middle nodes are substrings shared by several word forms, which means that if two word forms have different restrictions on the same substring, some of the restrictions may be violated. While the amount of violations was in practice very small, we ensured that no restrictions were violated in the end by applying the recursive algorithm only for the two first epochs, and then switching to Viterbi training.

In Viterbi training, each word is re-segmented to the most likely segmentation given the current model parameters using an extension of the Viterbi algorithm. We modified the implementation of Virpioja et al. (2013) to remove the previous segments of the word from the parameters before reanalyzing the word, and re-adding the segments of the new optimal segmentation afterwards. Additive smoothing with smoothing constant 1.0 was applied in the Viterbi search.

Prior to the Viterbi training, we flattened the tree structure so that the root nodes (word forms) link directly to the leaf nodes (morphs), thus removing any shared substrings nodes that are not actual morphs. This way all word forms are segmented independently and all the restrictions are followed.

2.1.3 Tuning the amount of segmentation

Omorfi-restricted Morfessor was tuned following Grönroos et al. (2015) to bring the number of tokens on the Finnish target side as close as possible to the English source side. The corpus weight hyper-parameter α was chosen by minimizing the sentence-level difference in token counts between the English and the segmented Finnish sides of the parallel corpus.

2.2 Rescoring n-best lists

Segmentation of the word forms increases the distances spanned by dependencies that should be modeled by the language model. To compensate for this, we apply a strong recurrent neural language model, TheanoLM. A recurrent language model is able to use arbitrarily long contexts without suffering from data sparsity, as opposed to n-gram language models, which are limited to a short context window. The additional language model is used in a separate rescoring step, to speed up translation, and for ease of implementation.

The TheanoLM model was trained on morphologically segmented data. Morphs occurring less than 1000 times in the full monolingual data were removed from the vocabulary, and replaced with the tag `<UNK>`. To create a class vocabulary, the morphs were embedded in a 300-dimensional space using word2vec (Mikolov et al., 2013). The embeddings were clustered into 2000 classes, using agglomerative clustering with cosine distance. Due to TheanoLM limitations, only the Europarl and News data (but not Common Crawl) were used for training.

The TheanoLM parameters were: 100 nodes in the projection layer, 300 LSTM nodes in the hidden layer, dropout rate 0.25, adam optimization with initial learning rate 0.01, and minibatch 16.

2.3 Morph boundary correction

One benefit of segmented translation is the ability to generate new compounds and inflections, that were not seen in the training data. However, the ability can also lead to errors, e.g when an English word frequently aligned to a compound modifier is translated using such a morph, even though there is no compound head to modify. The "dangling" morph boundary marker will then cause the space to be omitted, forming an incorrect compound with whatever word happens to follow.

For example, the Finnish pronoun *moni* (many) is also a frequent prefix, as in *monitoimi-* (multipurpose) or *monikulttuurinen* (multicultural). This resulted in an erroneous novel compound in *moniliberaalien keskuudessa* ("among the multiliberals"), which was corrected by introducing a space between *moni* and *liberaalien*, leading to a correct translation ("many among the liberals").

In the opposite type of error, compounds may be translated as separate words, or hyphenated compounds translated with the hyphen omitted.

We trained a neural network predictor to correct such errors by predicting the boundary type {space, empty, hyphen} as an additional post-processing step before joining the tokens.

The neural network takes as input both a token level representation, in the form of the same word2vec embeddings as used in rescoring, and a character level representation windowed to 4 characters before and after the boundary. The tokens are encoded by a bidirectional network of Gated Recurrent Units (Cho et al., 2014), while the characters are encoded by a feed-forward network.

Even though the boundary markers in the translation output are unreliable, they are a strong clue. Our predictor has access to the translated markers. During training markers were randomly corrupted to avoid relying too much on them.

2.4 Moses configuration

We used GIZA++ alignment. As decoding LMs, we used two SRILM n-gram models with modified-KN smoothing: a 3-gram and 5-gram model, trained from different data. Many Moses settings were left at their default values: phrase length 10, grow-diag-final-and alignment symmetrization, msd-bidirectional-fe reordering, and distortion limit 6.

The feature weights were tuned using MERT (Och, 2003), with BLEU (Papineni et al., 2002) of the post-processed hypothesis against a development set as the metric. 20 random restarts per MERT iteration were used, with iterations repeated until convergence.

The rescoring weights were tuned with a newly included script in Moses, which uses kb-MIRA instead of MERT.

3 Data

Our system participates in the constrained condition of the shared task. As parallel data, we used the Europarl-v8 and Wikititles corpora, resulting in 1 846 609 sentences after applying the Omorfi-restricted Morfessor segmentation and cleaning.

As monolingual data, we used the Finnish side of Europarl-v8, news.2014.fi.shuffled.v2, news.2015.fi.shuffled and Common Crawl. The total size of monolingual data after cleaning was 133 848 615 sentences, 2 135 919 860 morph tokens, and 11 771 367 morph types. Setting the frequency threshold to 1000 occurrences for the

	%BLEU, newstest		Example sentence
Configuration	2015	2016	Other applications could focus on muscle cells and insulin-producing cells, he added.
Omorfi-restricted Morfessor	10.77	11.27	Muissa sovelluksissa voi keskittyä lihas solujen ja insuliinia tuottavien solujen, hän lisäsi.
+boundary correction	10.83	11.27	Muissa sovelluksissa voi keskittyä lihassolujen ja insuliinia tuottavien solujen, hän lisäsi.
+rescoring	11.17	**11.73**	Muut sovellukset voivat keskittyä lihas soluja ja insuliinia tuottavia soluja, hän lisäsi.
+rescoring +boundary corr.	**11.21**	11.72	Muut sovellukset voivat keskittyä lihassoluja ja insuliinia tuottavia soluja, hän lisäsi.
Omorfi	10.00	10.59	Muut sovellukset voisi keskittyä lihassolujen ja insuliinia tuottavien soluja, hän lisäsi.
+boundary correction	10.07	10.61	Muut sovellukset voisi keskittyä lihassolujen ja insuliinia tuottavien soluja, hän lisäsi.
+rescoring	10.70	11.11	Muut sovellukset voivat keskittyä lihassoluja ja insuliinia tuottavien soluja, hän lisäsi.
+rescoring +boundary corr.	10.78	11.11	Muut sovellukset voivat keskittyä lihassoluja ja insuliinia tuottavien soluja, hän lisäsi.
Word baseline	10.48	10.65	Muut sovellukset voisivat keskittyä lihaksia ja insuliinia tuottavien solujen-, hän lisäsi.
Reference translation			Muut sovelluskohteet voisivat keskittyä lihassoluihin ja insuliinia tuottaviin soluihin, hän lisäsi.

Table 2: Results of automatic evaluation, in BLEU percentage points.

TheanoLM morph lexicon reduced the number of morph types to 121 735.

The complete monolingual data including the Common Crawl was only used for creating the morph lexicon and for training the 3-gram LM. For the 5-gram LM, the TheanoLM and the boundary predictor, the Common Crawl was omitted.

Because hyphenated compounds are much less frequent than non-hyphenated words, we enriched the training data for the boundary predictor by adding the list of words compounds containing a single hyphen and occurring more than 10 times in the full monolingual corpus.

4 Results

Results are summarized in Table 2, together with example translations produced by the different system configurations.

The Omorfi-restricted Morfessor segmentation leads consistently to an improvement over directly using the Omorfi segmentation. For all configurations on the newstest2016 set, and for newstest2015 without rescoring, the improvement is over +0.6 BLEU. On newstest2015 with rescoring, the improvement is slightly smaller, +0.47 BLEU.

Adding the TheanoLM rescoring increases BLEU between +0.4 and +0.7 BLEU. The increase is larger for the more aggressively segmented Omorfi system, supporting the conclusion that a strong language model is needed to compensate for the longer sequences.

In total, our best system results in a +1 BLEU improvement over the word baseline.

Boundary prediction gave a modest improvement of under +0.1 BLEU on the newstest2015 set, the effect on the newstest2016 set was neutral. While the predictor works reliably for the correct Finnish text it was trained on, manual inspection shows that the performance is erratic for disfluent translation output. Even while the minor cosmetic improvements are more common than errors, the benefit is hard to quantify.

Due to a mistake during data pre-processing, one of the n-gram language models penalizes the use of numbers. The problem affects all the evaluated systems and lowers the overall scores. However, it does not affect the increase in BLEU from the use of Omorfi-restricted Morfessor or rescoring. We verified this using BLEU of the test set with all source sentences containing numbers removed.

5 Conclusions

We propose a new morphological segmentation method, combining the strengths of rule-based and unsupervised morphology. We optimize the segmentation in a data-driven manner, aiming to balance granularity between the two languages, while restricting segmentation to a subset of the linguistic morph boundaries. Using this segmentation, we improve SMT quality over the linguistically accurate segmentation.

Using a neural morph boundary predictor to correct errors in the boundary markings does not lead to an improvement in BLEU.

In total, our best system results in a +1 BLEU improvement over the word baseline.

Acknowledgments

This research has been supported by the Academy of Finland under the Finnish Centre of Excellence Program 2012–2017 (grant n°251170), and LASTU Programme (grants n°256887 and 259934). Computer resources within the Aalto

University School of Science "Science-IT" project were used.

References

Fabienne Cap, Alexander M Fraser, Marion Weller, and Aoife Cahill. 2014. How to produce unseen teddy bears: Improved morphological processing of compounds in SMT. In *14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar.

Tagyoung Chung and Daniel Gildea. 2009. Unsupervised tokenization for machine translation. In *Empirical Methods in Natural Language Processing (EMNLP)*, Singapore. Association for Computational Linguistics.

Ann Clifton and Anoop Sarkar. 2011. Combining morpheme-based machine translation with post-processing morpheme prediction. In *ACL: HLT*, Portland, Oregon, USA, June. Association for Computational Linguistics.

Ryan Cotterell, Thomas Müller, Alexander Fraser, and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-markov models. In *CONLL*, Beijing, China. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *ACL-02 Workshop on Morphological and Phonological Learning*, Philadelphia, PA, USA, July. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1), January.

Adrià De Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. 2009. Minimum Bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of HLT-NAACL 2009: Short Papers*, Boulder, CO, USA. Association for Computational Linguistics.

Seppo Enarvi and Mikko Kurimo. 2016. TheanoLM - An Extensible Toolkit for Neural Network Language Modeling. *ArXiv e-prints*, May. `http://arxiv.org/abs/1605.00942`.

Mark Fishel and Harri Kirik. 2010. Linguistically motivated unsupervised segmentation for machine translation. In *Language Resources and Evaluation (LREC)*, Valletta, Malta.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2015. Tuning phrase-based segmented translation for a morphologically complex target language. In *Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September. Association for Computational Linguistics.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of HLT-NAACL*, New York, USA. 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, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *45th annual meeting of the ACL on interactive poster and demonstration sessions*, Prague, Czech Republic. Association for Computational Linguistics.

Minh-Thang Luong, Preslav Nakov, and Min-Yen Kan. 2010. A hybrid morpheme-word representation for machine translation of morphologically rich languages. In *Empirical Methods in Natural Language Processing (EMNLP)*, Cambridge, MA, USA. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems (NIPS)*, Lake Tahoe, NV, USA.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *41st Annual Meeting on Association for Computational Linguistics*, Sapporo, Japan. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *40th annual meeting of the association for computational linguistics*, Philadelphia, PA, USA. Association for Computational Linguistics.

Tommi A Pirinen, Antonio Toral, and Raphael Rubino. 2016. Rule-based and statistical morph segments in English-to-Finnish SMT. In *2nd International Workshop on Computational Linguistics for Uralic Languages*, Szeged, Hungary, Jan.

Tommi A Pirinen. 2015. Omorfi–Free and open source morphological lexical database for Finnish. In *20th Nordic Conference on Computational Linguistics (NODALIDA)*, Vilnius, Lithuania.

Raphael Rubino, Tommi Pirinen, Miquel Esplà-Gomis, Nikola Ljubešić, Sergio Ortiz Rojas, Vassilis Papavassiliou, Prokopis Prokopidis, and Antonio Toral. 2015. Abu-MaTran at WMT 2015 translation task: Morphological segmentation and web crawling. In *Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, September. Association for Computational Linguistics.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2014. Painless semi-supervised morphological segmentation using conditional random fields. In *European Chapter of the Association for Computational Linguistics (EACL)*, Gothenburg, Sweden, April. Association for Computational Linguistics.

Sara Stymne and Nicola Cancedda. 2011. Productive generation of compound words in statistical machine translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, Edinburgh, UK. Association for Computational Linguistics.

Sara Stymne, Nicola Cancedda, and Lars Ahrenberg. 2013. Generation of compound words in statistical machine translation into compounding languages. *Computational Linguistics*, 39(4).

Sami Virpioja, Jaakko J Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. *Machine Translation Summit XI*, 2007.

Sami Virpioja, Ville T. Turunen, Sebastian Spiegler, Oskar Kohonen, and Mikko Kurimo. 2011. Empirical comparison of evaluation methods for unsupervised learning of morphology. *Traitement Automatique des Langues*, 52(2).

Sami Virpioja, Peter Smit, Stig-Arne Grönroos, and Mikko Kurimo. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Report 25/2013 in Aalto University publication series SCIENCE + TECHNOLOGY, Department of Signal Processing and Acoustics, Aalto University, Helsinki, Finland.