# Multi-source synthetic treebank creation for improved cross-lingual dependency parsing

**Francis M. Tyers**
Department of Linguistics
Indiana University
Bloomington, IN
ftyers@prompsit.com

**Mariya Sheyanova**
School of Linguistics
Higher School of Economics
Moscow
marija.shejanova@gmail.com

**Alexandra Martynova**
School of Linguistics
Higher School of Economics
Moscow
alex250396@gmail.com

**Pavel Stepachev**
School of Linguistics
Higher School of Economics
Moscow
pavel.stepachev@yandex.ru

**Konstantin Vinogradovsky**
School of Linguistics
Higher School of Economics
Moscow
kvinog54@gmail.com

## Abstract

This paper describes a method of creating synthetic treebanks for cross-lingual dependency parsing using a combination of machine translation (including pivot translation), annotation projection and the spanning tree algorithm. Sentences are first automatically translated from a lesser-resourced language to a number of related highly-resourced languages, parsed and then the annotations are projected back to the lesser-resourced language, leading to multiple trees for each sentence from the lesser-resourced language. The final treebank is created by merging the possible trees into a graph and running the spanning tree algorithm to vote for the best tree for each sentence. We present experiments aimed at parsing Faroese using a combination of Danish, Swedish and Norwegian. In a similar experimental setup to the CoNLL 2018 shared task on dependency parsing we report state-of-the-art results on dependency parsing for Faroese using an off-the-shelf parser.

## 1 Introduction

In this paper, we describe and compare a number of approaches to cross-lingual parsing for Faroese, a Nordic language spoken by approximately 66,000 people on the Faroe Islands in the North Atlantic. Faroese is a moderately under-resourced language. It has a standardised orthography and fairly long written tradition, but lacks large syntactically-annotated corpora. There are however related well-resourced languages, such as Norwegian (both Bokmål and Nynorsk), Danish and Swedish for which large syntactically-annotated corpora exist.

Compared with the other Nordic languages, Faroese has a full nominal case system of four cases: Nominative, Genitive, Accusative and Dative, where the other languages have only a Genitive case. It has three grammatical genders, like Norwegian Nynorsk, but unlike Norwegian Bokmål, Danish and Swedish, which have a two-gender agreement system. Like the other Nordic languages, it is a verb-second (V2) language and the word order is generally similar. Faroese is however not mutually intelligible with any of the mainland Nordic languages.

Using these treebanks we perform experiments using two well-known methods, delexicalised parsing (Zeman and Resnik, 2008; McDonald et al., 2011) and synthetic treebanking using annotation projection (Tiedemann and Agić, 2016), and in addition propose a new method based on voting over possible projected trees using the maximum spanning tree algorithm. This can be thought of as creating a synthetic treebank where the tree for each sentence is the result of voting over the set of trees generated by parsing different translations.

The remainder of the paper is laid out as follows: Section 2 describes prior work on both Faroese and on cross-lingual dependency parsing; Section 3 describes the resources we used for the experiments, including a description of how the gold-standard for Faroese was made; Section 4 describes the methodology, including both the baseline models and our proposed method. Sections 5 and 6 de-



*Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150
Brussels, Belgium, November 1, 2018. 

scribe the experiments we performed and the results and discussion respectively and finally: Section 7 describes future avenues for research and Section 8 concludes.

## 2 Prior work

Our work is closely related to two main trends in cross-lingual dependency parsing. The first is multi-source delexicalised dependency parsing as described by McDonald et al. (2011).

The second is the work on synthetic treebanking by Tiedemann and Agić (2016); Tiedemann (2017). In these works, sentences in the target language (e.g. Faroese) is first translated by a machine translation system to a well-resourced language (e.g. Norwegian). The machine-translated Norwegian sentences are then parsed by a parser trained on a treebank of Norwegian, and word aligned to the Faroese originals. The output tree from the Norwegian parser is then *projected* back to the Faroese sentences via the word alignments.

In terms of voting for parse trees, the CoNLL shared task on dependency parsing in 2007 (Nivre et al., 2007) reported that using a similar architecture to the one we describe here, they were able to get significantly better results by combining the trees produced by the top three systems, and found that even after adding all the systems, including the worst-performing system, the performance did not drop below that of the top-performing system.

Our work is very similar to Agić et al. (2016), in that we use spanning tree to find the best parse in a graph that has been induced from aligned parallel corpora. However, their focus is on cross-linguality rather than on producing the best system for a related language, and as such the performance they report is lower.

It is also worth noting the work by Schlichtkrull and Søgaard (2017), who present a system that can learn from dependency graphs over tokens as opposed to over the well-formed dependency trees that are typically assumed for other systems.

In terms of dependency parsing specifically for Faroese, we can include the work by Antonsen et al. (2010), who apply a slightly-modified rule-based parser written for North Sámi to parsing Faroese. They achieved good results, F-score of over 0.98, on a small test set of 100 sentences. Unfortunately their work is not directly comparable as it relies on a very different annotation scheme to that which we use in our work, in addition they did not evaluate end-to-end results (the evaluation was done over gold standard POS and morphology).

## 3 Resources

In the experiments we used raw Faroese text extracted from Wikipedia, a manually created gold-standard corpus of trees, treebanks for the source languages (Danish, Swedish and Norwegian) and machine translation systems between the languages. The following subsections describe these resources.

### 3.1 Raw data

The Faroese raw data that we used in our experiments comes from Wikipedia dump which was preliminary cleaned of all the markup using the `WikiExtractor` script.[1] Then, both manually and via regular expressions, we deleted non-Faroese texts, poetic texts, reference lists, short sentences with little or no dependencies. All sentences containing only non-alphanumeric symbols were also deleted.

For sentence segmentation we used regular expressions splitting on sentence-final punctuation, but taking care to ignore month names, ordinal numbers and abbreviations. After cleaning the corpus we ended up with a total of 28,862 sentences. This data was used in the creation of the gold standard (§3.2) and in creating the parallel data used for the synthetic treebanking experiments (§4.2).

### 3.2 Gold standard

In order to evaluate the methods we needed to create a gold-standard treebank of Faroese. This was done manually by annotating sentences from the Faroese Wikipedia.[2] The gold standard contains 10,002 tokens in 1,208 sentences. The annotation procedure was as follows: We extracted sentences from the Faroese Wikipedia and analysed them using the Faroese morphological analyser and constraint grammar described by Trosterud (2009). This gave us a corpus where for each token in each sentence we had a lemma, a part of speech and a set of morphological features. These were checked manually and on top of these analyses, a dependency tree was added according to the guidelines in version 2.0 of Universal Dependencies (Nivre et al., 2016). Each tree was added manually by the

[1] https://github.com/attardi/wikiextractor

[2] The treebank is available online at https://github.com/UniversalDependencies/UD_Faroese-OFT.

| Treebank | Sentences | Tokens |
|---|---|---|
| UD_Swedish-Talbanken | 4,304 | 66,673 |
| UD_Danish | 4,384 | 80,378 |
| UD_Norwegian-Nynorsk | 14,175 | 245,330 |
| UD_Norwegian-Bokmaal | 15,696 | 243,887 |

**Table 1:** Number of sentences and tokens in UD treebanks for training the delexicalised models

first author in discussion with a native speaker of Faroese and members of the Universal Dependencies community.[3] The part-of-speech tags and features were converted automatically to ones compatible with Universal Dependencies using a lookup table and the longest-match set overlap procedure described in Gökırmak and Tyers (2017).

### 3.3 Other treebanks

For training the delexicalised models we used the following treebanks: UD_Swedish-Talbanken, UD_Danish-DDT (Johannsen et al., 2015), UD_Norwegian-Bokmaal (Øvrelid and Hohle, 2016) and UD_Norwegian-Nynorsk. Some statistics about these treebanks are presented in Table 1.

### 3.4 Machine translation

Faroese is not supported by the mainstream online machine translation engines and there are very few parallel sentence pairs available. For example, the widely-cited OPUS collection (Tiedemann, 2016) contains fewer than 7,000 sentences pairs for Faroese–Danish, Faroese–English, Faroese–Norwegian and Faroese–Swedish. This makes creating a corpus-based machine translation model unlikely to succeed. There is however a prototype rule-based machine translation system from Faroese to Norwegian Bokmål available through the Apertium project (Forcada et al., 2011).[4] This system has a vocabulary coverage of approximately 90% on the Faroese Wikipedia and supports translation of compound words. In addition to this system, systems for Norwegian Bokmål to Norwegian Nynorsk (Unhammer and Trosterud, 2009) and Norwegian Bokmål to Swedish and Danish also exist. As a result of this, we decided to use pivotting via Norwegian Bokmål to produce the translations (see §4.2).

[3] Some of the discussions can be found in the issues page of the UD_Faroese-OFT repository: https://github.com/UniversalDependencies/UD_Faroese-OFT/issues

[4] https://github.com/apertium/apertium-fao-nor

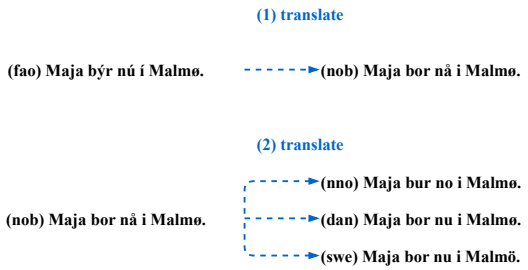


**Figure 1:** Example of pivot translation from Faroese to Swedish, Danish and Norwegian Nynorsk via Norwegian Bokmål. The sentence *Maja býr nú í Malmø* translates in English as 'Maja now lives in Malmø'. The translation to the other Nordic languages is word-by-word and monotonic.

## 4 Methodology

In this section we describe the two baseline methods and our multi-source approach.

### 4.1 Delexicalised parsing

For the delexicalised parsing baseline, we trained delexicalised models on the Swedish, Danish, Norwegian Bokmål and Norwegian Nynorsk Universal Dependencies treebanks. Delexicalised models are models trained only on the sequence of POS-tags and morphological features, omitting both lemmas and surface forms. The idea behind this is to make the model maximally language independent.

### 4.2 Annotation projection

For each of the source languages (Swedish, Danish, Norwegian Bokmål, and Norwegian Nynorsk), we first translated the Faroese Wikipedia (§3.1) to that language using the Apertium machine translation system. In the case of Swedish, Danish and Norwegian Nynorsk, the translation is pivoted via Norwegian Bokmål. This is demonstrated in Figure 1.

The original Faroese text and the translation is then aligned using fastalign (Dyer et al., 2013), a word-aligned based on IBM Model 2. Both translations and alignments are largely word-for-word and monotonic.

We then parse the translation using a lexicalised model trained on the training portion of the relevant treebank using UDPipe (Straka and Straková, 2017). This results in a collection of: original Faroese sentences, translations of those sentences, a word-by-word alignment between the Faroese sentences and the translated sentences, and a tree for each of the translated sentences.

The next step is to take the trees over the translated sentences and *project* them back to the original Faroese sentences, as is shown in Figure 2.

The final trees are then used for training a *lexicalised* model using UDPipe for parsing Faroese.

| Language | Sentences | Tokens |
|---|---|---|
| Swedish | 28,701 | 758,999 |
| Danish | 28,632 | 768,662 |
| Norwegian Bokmål | 28,016 | 765,203 |
| Norwegian Nynorsk | 28,611 | 753,597 |

**Table 2:** Number of valid sentences in synthetic UD treebanks for single-language models

### 4.3 Multi-source projection

With multi-source projection we add some additional steps. Instead of training a model on sentences which have had annotation projected from a single source language, we take into account the annotation for the sentence from all of the languages.

We first build a dependency graph for each Faroese sentence using the arcs found in all of the parsed translations of that sentence. The arcs are weighted, like in the voting scheme from the CoNLL-07 shared task (Nivre et al., 2007), such that each language is counted as a single vote for that arc. The dependency relations are voted for independently after the best tree has been found.

To find the best tree in the weighted graph, we use the maximum-spanning tree (MST) algorithm of Chu (1965); Edmonds (1967). This algorithm is widely used in dependency parsing, cf. McDonald et al. (2005). The algorithm is composed of the following steps:

1. For each vertex, pick the the incoming edge with the highest weight.

2. Check the graph for cycles. If there are no cycles and the graph is a tree, then return this graph as the resulting MST.

3. If there are cycles, then, for each cycle, isolate the cycle from the tree, find the incoming (to any vertex of the cycle), edge with the highest weight, then remove all the edges within the cycle which conflict with it.

4. Then repeat the steps 2-3 until there are no cycles.

Figure 3 shows the graph produced from the running example and the result of running the spanning tree algorithm.

## 5 Experiments

In order to evaluate the performance of multi-source synthetic treebank model, we conduct several experiments in which we compare the performance of our models to the baseline methods: delexicalised parsing (see §4.1) and synthetic treebanking (see §4.2).

For each model, we trained tagger and parser UDPipe models with the default settings (20 epochs for tagger and 10 epochs for parser).

## 6 Results

Here we present a comparison between the performance of baseline models described in Section 4 and that of the model trained on the multi-source synthetic treebank described in Section 4.3 measured against the gold standard.

Table 3 shows the F-measure for POS tagging and labelled-attachment score (LAS) and unlabelled-attachment score (UAS) for dependency relations. The best results for each approach are shown in bold.

## 7 Future work

One promising avenue for future work is to improve the way in which trees are projected into the lesser-resourced language. At the moment this is done in a deterministic fashion using the 1-best alignment from the word aligner. This has two primary drawbacks: (1) It could be however that there exist better alignments, but we miss them by choosing only the best; and (2) we then have to use imperfect heuristics to attempt to make a valid tree when the alignments do not result in one. One idea we have had would be to view the projection problem as one of finding the best tree in a graph of alignments. These alignments could come from several word aligners, or even from using simple attachment rules such as in e.g. Alonso et al. (2017).

Another avenue is to improve how arcs in the projected graph are weighted. At the moment we do only raw voting, but information from other languages in terms of distribution of part-of-speech tags, features and dependency relations could potentially improve the results.

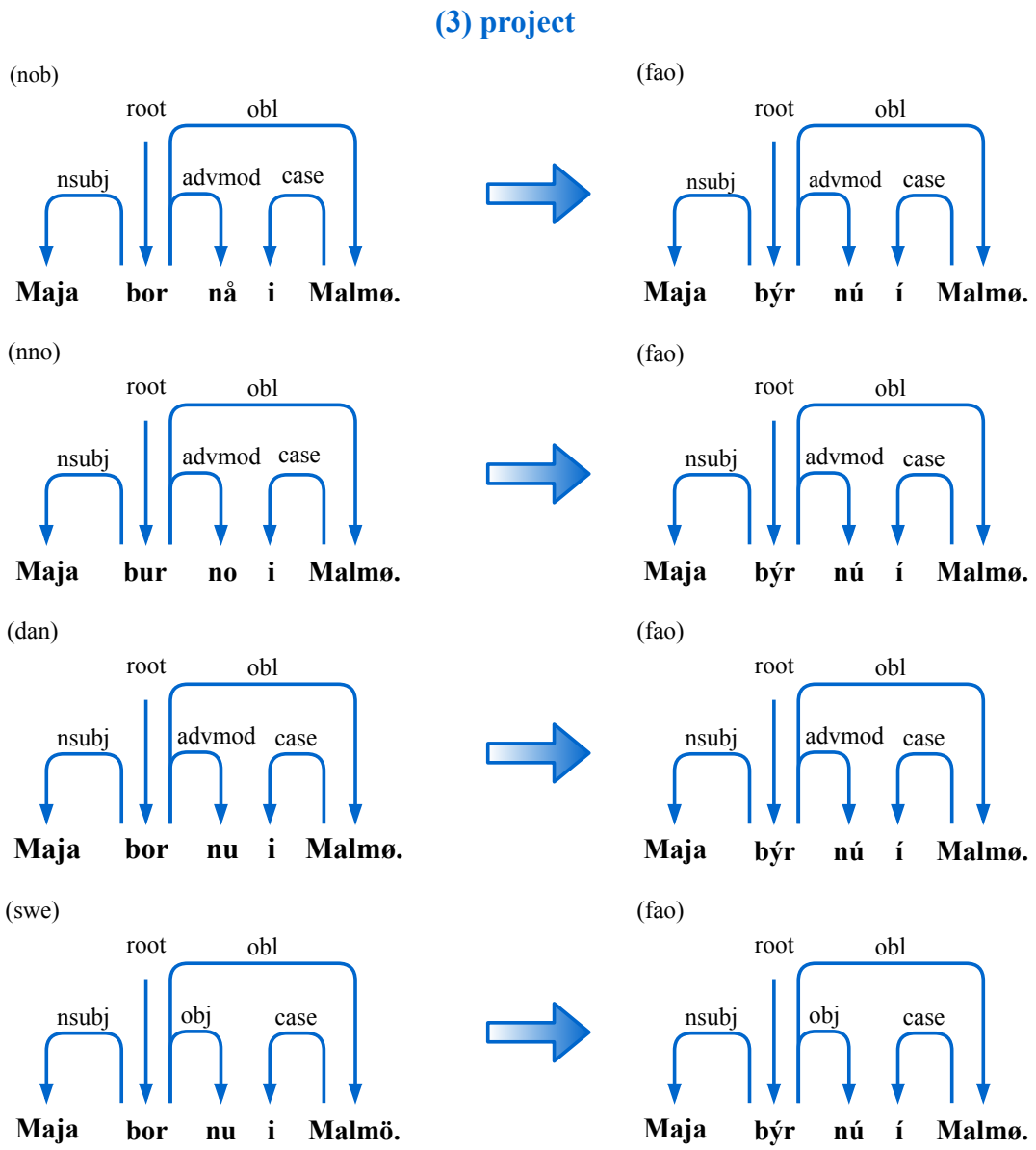


**Figure 2:** The sentences in the source languages are parsed, and then the trees are projected via the alignments back to the target language. In this case, the trees are identical with the exception of the annotation of *nu* 'now' as an object `obj` in Swedish.

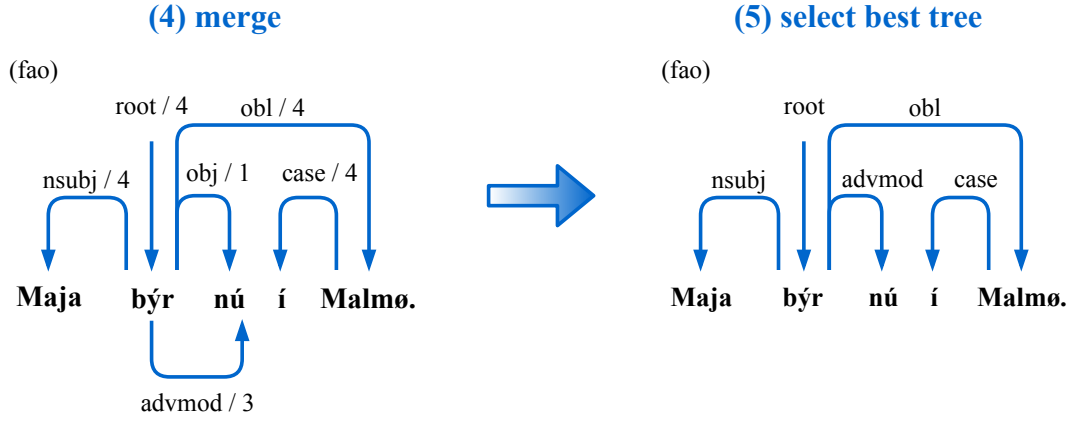


**Figure 3:** The projected trees from Figure 2 are merged into a weighted dependency graph where the weight of each edge is the number of times that edge is seen in the source trees. After merging the spanning tree algorithm is run to find the tree with the highest weight.

| Model | Delexicalised | | | Projected | | |
|---|---|---|---|---|---|---|
| | POS | UAS | LAS | POS | UAS | LAS |
| Swedish | 43.83 | 23.14 | 10.32 | 73.06 | 65.66 | 58.53 |
| Danish | 46.15 | 21.27 | 13.01 | 74.76 | 68.74 | 59.84 |
| Norwegian Bokmål | 44.29 | 24.51 | 15.62 | **74.89** | **72.04** | **63.95** |
| Norwegian Nynorsk | **51.30** | **27.76** | **18.93** | 72.93 | 70.62 | 62.27 |
| Multi-source | — | — | — | **74.49** | **72.90** | **64.43** |

**Table 3:** Results for the systems. Delexicalised models are trained directly on the target language treebank and applied directly. In bold are the best results for delexicalised parsing, projected parsing and parsing with multi-source trees. In all cases the multi-source model outperforms all others.

In addition, we would like to try increasing the number of trees used to build the graph in the source language. One possibility is to use different parsers to generate different trees, and another is to use more machine translation systems to produce more translations to align.

We would also like to try the approach with other language groups for which there are several related treebanks in the Universal Dependencies project, for instance Upper Sorbian.

## 8 Conclusion

We have presented a method of creating synthetic training data for parsing a moderately under-resourced language for dependency parsing by using pivot machine translation into several closely-related better-resourced languages. By training an off-the-shelf parser on this synthetic treebank we are able to substantially improve on the state of the art for dependency parsing of Faroese, a moderately under-resourced language. All of the code is available under a free/open-source licence online.[5]

## Acknowledgements

The article was prepared within the framework of the Academic Fund Programme at the National Research University Higher School of Economics (HSE) in 2016 — 2018 (grant №17-05-0043) and by the Russian Academic Excellence Project «5-100». The authors would like to thank Bjartur Mortensen for his help in preparing the gold standard, and the anonymous reviewers for their helpful comments.

[5] https://github.com/ftyers/cross-lingual-parsing

## References

Agić, □., Johannsen, A., Plank, B., Martínez Alonso, H., Schluter, N., and Søgaard, A. (2016). Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Alonso, H. M., Željko Agić, Plank, B., and Søgaard, A. (2017). Parsing Universal Dependencies without training. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 230–240.

Antonsen, L., Trosterud, T., and Wiechetek, L. (2010). Reusing grammatical resources for new languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC10*, pages 2782–2789.

Chu, Y.-J. (1965). On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of NAACL-HLT 2013*, pages 644–648. Association for Computational Linguistics.

Edmonds, J. (1967). Optimum branchings. *Journal of Research of the National Bureau of Standards*, 71:233–240.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Gökırmak, M. and Tyers, F. M. (2017). A dependency treebank for Kurmanji Kurdish. In *Proceedings of the Fourth International Conference on Dependency Linguistics (DepLing, 2017)*, pages 64–73.

Johannsen, A., Martínez Alonso, H., and Plank, B. (2015). Universal dependencies for danish. In *Proceedings of the 14th International Workshop on Treebanks and Linguistic Theories (TLT14)*.

McDonald, R., Pereira, F., Ribarov, K., and Hajič, J. (2005). Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.

McDonald, R., Petrov, S., and Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 62–72.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC16*.

Nivre, J., Hall, J., Kübler, S., McDonald, R. T., Nilsson, J., Riedel, S., and Yuret, D. (2007). The conll 2007 shared task on dependency parsing. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 915–932.

Schlichtkrull, M. S. and Søgaard, A. (2017). Cross-lingual dependency parsing with late decoding for truly low-resource languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 222–229. Association for Computational Linguistics.

Straka, M. and Straková, J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Tiedemann, J. (2016). Opus - Parallel corpora for everyone. *Baltic Journal of Modern Computing*, 4(2).

Tiedemann, J. (2017). Cross-lingual dependency parsing for closely related languages – Helsinki's submission to VarDial 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 131–136. Association for Computational Linguistics.

Tiedemann, J. and Agić, Z. (2016). Synthetic treebanking for cross-lingual dependency parsing. *Journal of Artificial Intelligence Research*, 55:209–248.

Trosterud, T. (2009). A constraint grammar for Faroese. In *Proceedings of the NODALIDA 2009 workshop Constraint Grammar and robust parsing*.

Unhammer, K. and Trosterud, T. (2009). Reuse of free resources in machine translation between Nynorsk and Bokmål. In *Proceedings of the First International Workshop on Free/Open-Source Rule-Based Machine Translation*, pages 35–42.

Zeman, D. and Resnik, P. (2008). Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 35–42.

Øvrelid, L. and Hohle, P. (2016). Universal Dependencies for Norwegian. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC16*.