Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4827–4832
Marseille, 20-25 June 2022

The Norwegian Dialect Corpus Treebank

Andre Kåsen[†], Kristin Hagen[*], Anders Nøklestad[*], Joel Priestley[*], Per Erik Solberg[†] and Dag Trygve Truslew Haug[*]
[†]National Library of Norway
{andre.kasen, per.solberg}@nb.no

[*]Department of Linguistics and Scandinavian Studies
{kristiha, noklesta, joeljp, daghaug}@uio.no

Abstract

This paper presents the NDC Treebank of spoken Norwegian dialects in the Bokmål variety of Norwegian. It consists of dialect recordings made between 2006 and 2012 which have been digitised, segmented, transcribed and subsequently annotated with morphological and syntactic analysis. The nature of the spoken data gives rise to various challenges both in segmentation and annotation. We follow earlier efforts for Norwegian, in particular the LIA Treebank of spoken dialects transcribed in the Nynorsk variety of Norwegian, in the annotation principles to ensure interusability of the resources. We have developed a spoken language parser on the basis of the annotated material and report on its accuracy both on a test set across the dialects and by holding out single dialects.

Keywords: treebanks, spoken language, dialects, Norwegian, spoken language parser

1. Introduction

In this article, we present the Norwegian Dialect Corpus Treebank – a treebank of spoken Norwegian dialects transcribed in the Bokmål variety of Norwegian from The Nordic Dialect corpus (NDC; Johannessen et al. (2009)). The project has been carried out within the CLARINO+ project,[1] a Norwegian project in the pan-European CLARIN infrastructure.

In this chapter we will start to introduce the NDC Treebank and relate our work to the two other dependency treebanks for spoken and written Norwegian. In chapter 2 we will go in to different aspects of the morphosyntactic annotation in more detail and describe how the work was undertaken. In chapter 3 we report on experiments with training a spoken language parser on the resulting treebank. Chapter 4 describes the accessibility of the treebank in the search interface Glossa and how the treebank can be downloaded.

The NDC Treebank consists of 4587 speech segments, overall 66009 tokens, from 17 different Norwegian dialects from south, west, east and north of Norway. The geographical distribution of the dialects is visualized in Figure 1 below. The recordings in The Nordic Dialect Corpus were made between 2006 and 2012 and comprise both interviews and more informal conversations between pairs of speakers. The transcriptions are therefore filled with spoken language phenomena, such as overlaps, pauses and various types of disfluencies. The NDC Treebank is annotated with morphological information and dependency-style syntactic analysis.

The NDC Treebank project is related to the two other dependency treebanks made for Norwegian: The Norwegian Dependency Treebank (NDT; Solberg et al. (2014)) with mostly written texts and The LIA Treebank of Spoken Norwegian Dialects (Øvrelid et al., 2018).

The NDC Treebank differs from the other spoken language treebank, LIA, in two major ways. First, the LIA Treebank is based on older dialects, with recordings made between 1950 and 1990, while the recordings in NDC are from between 2006 and 2012 as mentioned above. Second and most important, the transcriptions in LIA are written in the Norwegian standard Nynorsk, while the NDC transcriptions are written in Bokmål. The NDT Treebank contains written texts both in Bokmål and Nynorsk.[2]

Although the LIA Treebank and the NDT treebank have become important sources for both Norwegian language research and language technology, there was a need for a treebank for spoken dialects transcribed to Bokmål. Bokmål is the dominant written standard in Norway and it closely resembles the spoken language in the Oslo area. Many of the available Norwegian spoken language corpora[3] are transcribed in Bokmål. With a Bokmål treebank of spoken dialects, we can train taggers and parsers to annotate these corpora.

In the NDC Treebank project we have reused the annotation guidelines from the earlier two projects and also tried to utilize the experiences gained from them both. We also used data from the earlier treebanks when training parsers in this NDC project. Finally, for

[1] https://clarin.w.uib.no/about/

[2] Norway has two written standards for Norwegian, Bokmål which originates from Danish and Nynorsk which was constructed from Norwegian dialects and made an official standard in 1885. The standards are mutually understandable.

[3] See an overview of spoken corpora at the Textlaboratory here: https://www.hf.uio.no/iln/english/about/organization/text-laboratory/services/index.html#speech

the NDC Treebank, we chose transcriptions from the same areas as the transcriptions in the LIA Treebank. This choice opens up for interesting comparisons of the two written standards.

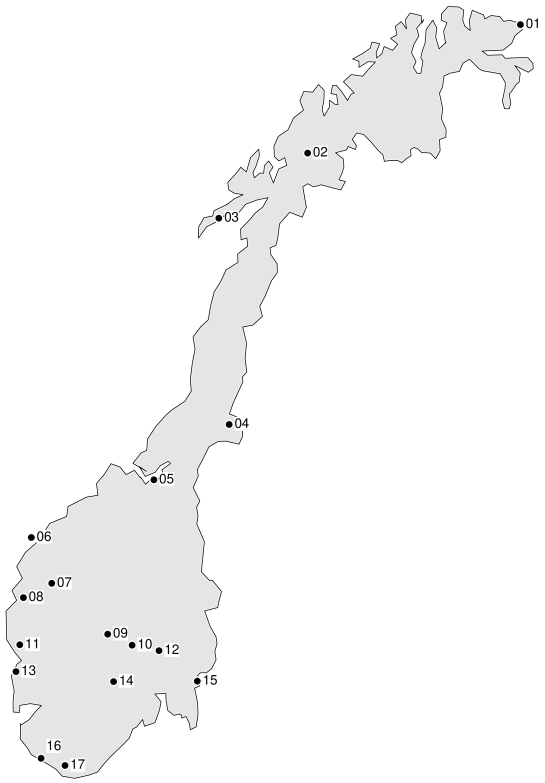

Figure 1: Geographical distribution of the dialects in the treebank. From top to bottom it list as follows: 01) Vardø, 02) Kirkesdalen, 03) Stamsund, 04) Lierne, 05) Trondheim, 06) Herøy, 07) Jølster, 08) Hyllestad, 09) Ål, 10) Flå, 11) Bergen, 12) Jevnaker, 13) Bømlo, 14) Hjartdal, 15) Rømskog, 16) Sokndal and 17) Lyngdal.

2. Morphosyntactic Annotation

2.1. Transcription and Segmentation

In the Nordic dialect corpus, there are two transcriptions of each recording, one phonetic-like and one orthographic. For both, only a few special characters are in use except for the Norwegian letters *æ*, *ø*, *å* and some characters used to indicate pauses, tags for laughter, coughing etc. Only the orthographic version is used in the treebank. For more details, see Øvrelid et al. (2018).

The transcriptions in the NDC Treebank are divided into speech segments. The segments often correspond to a written sentence. However, characteristics of spoken language like disfluencies, repetitions and incomplete sentences give a lot of exceptions. Many segments in the NDC Treebank can also be long and complicated because the segmentation was originally done for building the Nordic Dialect Corpus and was not only based on syntactic principles but also on content criterions to give meaningful search results in the corpus. We decided to keep the original segments for two reasons. First, we wanted a parser trained on the NDC Treebank to be able to parse the rest of the corpus and other spoken corpora successfully, and the long segments will provide realistic training material. Second, we wanted to keep the original segments and their associated time codes from the Nordic Dialect corpus to make a searchable treebank in Glossa with corresponding audio and video, see section 4.

2.2. The annotation principles

Both the morphological and syntactic annotation in the NDC Treebank follow the LIA Treebank (Øvrelid et al., 2018), which extends the annotation scheme of the Norwegian Dependency Treebank (*NDT*; Solberg et al. (2014)) with a treatment of spoken-language phenomena. While there are versions of both the LIA Treebank and the NDT in the Universal Dependencies annotation standard (Øvrelid and Hohle, 2016; Øvrelid et al., 2018), we have chosen the original annotation standard of those treebanks for NDC, as there exist detailed annotation guidelines which could be reused (Kinn et al., 2014). According to the annotation study in Skjærholt (2014), the NDT had an inter-annotator agreement of 98%, among the best scores in the study, which is a good indicator of the quality of the annotation guidelines. An updated and public available conversion procedure is in the making. It will be based on the work in Øvrelid and Hohle (2016) and written in the the GREW tool (Guillaume, 2021).

While the Universal Dependencies standard favors lexical words as heads, the NDT/LIA standard has a hybrid approach where some constructions have lexical heads, while others have functional heads. Function words which must be present, are taken as heads. Prepositions belong to this group, taking the prepositional complement as a dependent. Similarly, a finite clause is always headed by the finite verb, even when this verb is an auxiliary. Lexical, non-finite verbs are dependents of the finite auxiliary. Verbal arguments and modifiers, in turn, are dependents of the lexical verbs, except for subjects, which are dependents of the finite auxiliary.

Function words which are not always present or can be dropped, are not heads. In that way, constructions are analyzed in the same manner, regardless of the presence or absence of function words. Nouns are heads in nominal constructions, taking determiners as dependents, as many nominal constructions lack determiners. In a similar vein, complementizers are dependents of verbs, as complementizers are frequently dropped. In the case of coordination, the first conjunct is the head, taking the subsequent conjuncts as dependents with a dedicated label. Conjunctions are dependents of the closest conjunct to the right.

The annotation scheme aims at being as linguistically accurate as possible, following the Norwegian Refer-

ence Grammar (Faarlund et al., 1997). At the same time, annotation speed and accuracy are given high priority as well. Due to this, there is no distinction between different types of adverbials or selected and modifying adverbials. All receive the same dependency relation *ADV*. We refer to Solberg et al. (2014, 789–792) for more details on these annotation choices and the motivation behind them.

As for spoken language syntax, we follow the LIA guidelines. The text contains a number of extra-linguistic tokens: some of these, such as pauses, are integrated in the syntactic tree as dependents of the following word. This will facilitate later use of the corpus to study e.g. the relationship between prosodic breaks and syntactic constituency. Other extra-linguistic tokens, indicating e.g. laughter, sighs etc. are simply ignored. Another frequent phenomenon is disfluency, where we distinguish repairs (structures where a false start is subsequently repaired) and restarts (segments that are not completed nor repaired), annotated with REP and SLETT ('deletion' in Norwegian). Both types are very common in the corpus, but the distinction is in many cases subtle. Finally, there are phenomena that are found also in written language but which are much more common in spoken language, such as ellipsis and discourse particles. For more details on how we deal with spoken language syntax, see Øvrelid et al. (2018, section 4.2).

Figure 2 shows a sample annotated sentence. The hash symbols represent pauses. We see that between the two pauses, we have (in this case) a well-formed sentence conforming to the written standard and analyzable in terms of the standard relations of the NDT. To the left and to the right, separated by the pauses, there are fragments that are internally analyzable (to some extent), but not complete. They are attached to the main verb via REP and SLETT. The annotator has chosen to regard the main sentence ('he knew exactly…') as a repair of the initial fragment ('it was indeed…'), but this is clearly debatable.

Notice that there are non-trivial interactions between segmentation and syntactic annotation. For example, in Figure 2, the final fragment after the break is attached with a long rightwards SLETT edge. If it had instead been segmented with the following utterance, it would be attached with a leftwards edge that would likely be much shorter (because the main verb is in second position in Norwegian). The actual segmentation also forces the label SLETT: there is no following material so the fragment cannot possibly be considered a repair structure. If the fragment was part of the following segment, REP might have been an option. Finally, an alternative segmentation would take the final fragment as a separate segment, in which case it would be a root with the FRAG relation.

2.3. Morphosyntactic Preprocessing and Manual Correction

In the NDC Treebank project we could use the already existing Norwegian treebanks, LIA and NDT, for training preprocessing tools. Ideally, such a tool should be available as a single pipeline in a library like spaCy.[4] This is not yet the case, but an important point for future work. In the present work a new composite pipeline was established. The lemmatization was carried out in NorLem,[5] while morphosyntactic features and part-of-speech tags were assigned with custom trained models in spaCy based on a transfer from NDT to LIA. The resulting files were parsed with models similar to the ones described in Kåsen (2020).

In the final step, linguistically trained annotators corrected the output of the morphosyntactic preprocessing using the conllu editor.[6] This was done by two student annotators. The annotators consulted with each other about their choices in difficult constructions, and met every week with two senior members of the project for further discussions. In the end most sentences were handled twice, either proof-read by the other student or one of the seniors. Interannotator agreement was not measured.

Both dependency relation, syntactic labels, lemmatization and part-of-speech tags were corrected in conllu editor. In total, 16001 of the tokens needed to be edited. This represents 24.2% of the total number of tokens. If the quality of the preprocessing was close to the parsing accuracies that we report in section 3, this seems reasonable. With the tool MaltEval (Nilsson and Nivre, 2008), we find that the most frequent correction done by the annotator is from a formal subject (FSUBJ) to subject (SUBJ) and subject predicative (SPRED) to potential subject (PSUBJ). This is most likely due to the fact that in Norwegian expletive constructions are quite frequent. Furthermore there is severe types of expletive constructions that might be hard to differentiate. See Bouma et al. (2018) for a review of expletives and treebanks.

Treebank Combination	**LAS**	**UAS**
NDT	49.98	60.07
NDC	76.18	83.41
NDT$_{nob}$+ NDC	77.87	84.25
NDT + NDC	78.52	**85.04**
NDT + LIA + NDC	**78.61**	84.84

Table 1: Scores for the overall treebank embedding experiments on the NDC test split

[4] https://spacy.io/

[5] https://github.com/emanlapponi/norlem-norwegian-lemmatizer

[6] https://github.com/Orange-OpenSource/conllueditor

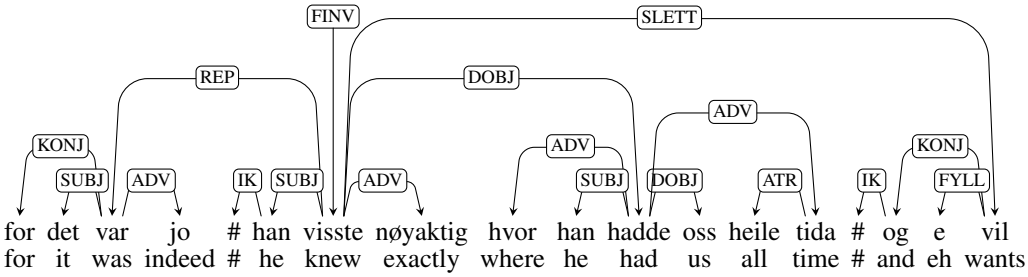

Figure 2: Example of manually annotated dependency graph with spoken language specific relations from the NDC test set.

Dialect Area	LAS	UAS
Ål	83.45	88.26
Bergen	82.84	87.73
Bømlo	81.90	86.44
Flå	83.79	88.08
Herøy	83.33	88.59
Hjartdal	84.65	88.82
Hyllestad	84.82	89.21
Jevnaker	80.16	85.08
Jølster	81.79	85.82
Kirkesdalen	79.41	86.20
Lierne	87.10	91.10
Lyngdal	84.24	88.56
Rømskog	85.43	89.53
Sokndal	84.30	88.89
Stamsund	84.35	88.95
Trondheim	**87.81**	**91.56**
Vardø	84.19	88.52

Table 2: Scores for the dialect-wise evaluation.

3. Spoken language parsers

3.1. Parser

UUParser is a transition-based dependency parser as described in de Lhoneux et al. (2017) and builds on the BISTParser found in Kiperwasser and Goldberg (2016). It also supports model training with more than one treebank via treebank embeddings (Stymne et al., 2018) or dataset embeddings (van der Goot and de Lhoneux, 2021). In essence, the treebank embeddings should encode similarities between the treebanks, in our case also between modalities (speech and writing), and still maximize the parser performance for each treebank alone (van der Goot and de Lhoneux, 2021, p. 22). An important assumption for the present work is that both written and spoken language share a core of grammatical relations that can be represented in a dependency grammar framework, but this is not a trivial issue (see Miller and Weinert (1998) or Ortmann and Dipper (2019)). That being said, Dobrovoljc and Martinc (2018) quantifies that a mixing of modalities i.e. spoken and written language can actually lower parser performance. However, Dobrovoljc and Martinc (2018) does not use treebank embeddings. And moreover, Stymne et al. (2018, p. 619)'s approach "has the advantage of producing a single flexible model for each language, regardless of the number of treebanks." They also note with reference to Velldal et al. (2017) that a concatenation of treebanks for Norwegian is dependent on machine translation in order to achieve best performance.

3.2. Setup

The treebank was split with the UD guidelines for dataset release[7] in mind. Care was taken to ensure an as even as possible distribution of each dialect between the splits, thus the dev and test folds are just under 10K. Then, several parser models were trained with UUParser.

In large, two kinds of experiments were conducted: 1) in the style of Stymne et al. (2018) with all the different treebanks for Norwegian, and 2) a cross validation inspired evaluation where every single dialect in NDC served as a test set. The latter experiment represents a realistic use case for a dialect treebank, as we may want to use the parser on material from a dialect that is not represented in the current corpus. It also provides interesting information about what dialects are most easy and difficult to parse.

3.3. Results

Tables 1 and 2 show the labelled attachment score (LAS) and unlabelled attachment score (UAS) computed for the different experiments on the test set of the NDC Treebank with the same treebank as a proxy treebank.[8]

Row 1 in Table 1 lists the scores for predicting with only NDT (i.e. NDT split in a Bokmål and Nynorsk file) whereas rows 3 and 4 list training with only the

[7] *If you have between 30K and 100K words, take 10K as test data, 10K as dev data and the rest as training data.*

[8] A proxy in this context is a way to signal or prompt the parser model the type of input it can expect. Of course if the input is unseen, i.e. not belonging to any of the data points in the training set, deciding what proxy is optimal is not a trivial issue.

Bokmål fragment together with the NDC Treebank and both Bokmål and Nynorsk in addition to the NDC Treebank. The last row is the result for training with all available treebanks for Norwegian. The inclusion of the LIA Treebank also gives an extra boost in performance. And even though the NDC Treebank only contains Bokmål transcripts, the Nynorsk part of NDT and of LIA seem to contribute in a positive manner.

An important point is that spoken language data is a necessary addition in order to reach an acceptable level of performance. When only parsing spoken language data with written language (here with the Bokmål part as a proxy), the performance is poor. But training with only the NDC Treebank with the scores given in the second row in Table 1 yields an acceptable score. However, the performance is improving as more relevant data is added.

For the dialect-wise evaluation in Table 2, the results are overall better than for the experiments in Table 1. Even the worst results in Table 2 are better than the best result from Table 1. One of the most salient factors might be that there is more training data available in this setup. The performance range across the dialects is quite big, but there can be many reasons for this, including variation among the dialect informants in the use of dysfluencies and other spoken features that are hard to parse, so it is hard to argue that the differences are due to syntactic variation between dialects. In fact, the consistently good numbers in Table 2 suggest that the variation is small enough to allow for parsing across dialects. This is useful because a future use case for the parser will be to analyze new material in dialects that are not covered in the training data.

4. Accessibility

The treebank is made available for search in Glossa, a web-based search interface for written and spoken mono- and multilingual corpora. Glossa enables the user to formulate linguistic queries, potentially restricting what part of the corpus is queried through metadata selection. The results are presented as Keyword-In-Context (KWIC) concordances and statistical summaries in the form of frequency lists, metadata distributions etc. For many spoken corpora, the search results are accompanied by audio and video clips, on-demand spectrographic analysis and maps showing the geographical distribution of dialect forms.

Glossa is designed with a large focus on user-friendliness. It allows the user to choose between three different search interfaces with varying power and ease-of-use. The first is a simple text input similar to what is found in web search engines such as Google or Bing, where the user can search for a word form or a phrase, potentially with truncation. The second one contains a sophisticated set of text inputs, checkboxes, buttons and drop-down menus that enables the user to formulate advanced queries using an intuitive graphical interface. The third variant allows the user to formulate queries directly in the query language of the underlying search engine (The IMS Open Corpus Workbench; (Evert and Hardie, 2011)), and take advantage of its capabilities to the fullest extent.

In keeping with the philosophy that user-friendliness is paramount, the options for searching the NDC Treebank in Glossa are restricted to those that can be presented in an intuitive graphical interface. In the initial version at least, this amounts to searching only for dependent labels and not for the head of a relation. On the other hand, syntactic search can be combined with search for morphosyntactic information, lemma or phonetic form, making it possible to formulate queries such as *Subject preceded by an adjective* or *Utterance final singular pronominal Objects*.

To enable users to perform more advanced searches in dependency structures, we intend to get the treebank integrated into the INESS search system (Rosén et al., 2012), which has been developed at the University of Bergen and which already includes the LIA Treebank previously developed by our team. This system enables very powerful syntactic searches, but using it requires a lot more technical skills. Offering the NDC Treebank in both systems will give users a wide range of options for search. Finally, the treebank can also be downloaded in CONLL format, allowing users to apply their own processing to the data.[9]

5. Acknowledgements

This project was supported by the Research Council of Norway under grant no. 295700 CLARINO+.

6. Bibliographical References

Bouma, G., Hajic, J., Haug, D., Nivre, J., Solberg, P. E., and Øvrelid, L. (2018). Expletives in universal dependency treebanks. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 18–26.

de Lhoneux, M., Stymne, S., and Nivre, J. (2017). Arc-hybrid non-projective dependency parsing with a static-dynamic oracle. In *Proceedings of the The 15th International Conference on Parsing Technologies (IWPT).*, Pisa, Italy.

Dobrovoljc, K. and Martinc, M. (2018). Er... well, it matters, right? on the role of data representations in spoken language dependency parsing. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 37–46.

Evert, S. and Hardie, A. (2011). Twenty-first century corpus workbench: Updating a query architecture for the new millennium.

Faarlund, J. T., Lie, S., and Vannebo, K. I. (1997). *Norsk referansegrammatikk*. Universitetsforlaget.

Guillaume, B. (2021). Graph matching and graph rewriting: Grew tools for corpus exploration, maintenance and conversion. In *Proceedings of the 16th*

[9] https://github.com/textlab/spoken_norwegian_resources

Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, pages 168–175.

Johannessen, J. B., Priestley, J., Hagen, K., Åfarli, T. A., and Vangsnes, Ø. A. (2009). The nordic dialect corpus–an advanced research tool. In *Proceedings of the 17th Nordic Conference of Computational Linguistics (NODALIDA 2009)*, pages 73–80.

Kåsen, A. (2020). Mot en trebank for amerikanorsk. *Oslo Studies in Language*, 11(2):243–250.

Kinn, K., Solberg, P. E., and Eriksen, P. K. (2014). NDT guidelines for morphological and syntactic annotation. Technical report, National Library of Norway, Oslo.

Kiperwasser, E. and Goldberg, Y. (2016). Simple and accurate dependency parsing using bidirectional LSTM feature representations. *Transactions of the Association for Computational Linguistics*, 4:313–327.

Miller, J. and Weinert, R. (1998). *Spontaneous spoken language: Syntax and discourse*. Clarendon Press.

Nilsson, J. and Nivre, J. (2008). Malteval: an evaluation and visualization tool for dependency parsing. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Ortmann, K. and Dipper, S. (2019). Variation between different discourse types: Literate vs. oral. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 64–79.

Øvrelid, L. and Hohle, P. (2016). Universal Dependencies for Norwegian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1579–1585, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Øvrelid, L., Kåsen, A., Hagen, K., Nøklestad, A., Solberg, P. E., and Johannessen, J. B. (2018). The LIA treebank of spoken Norwegian dialects. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Rosén, V., De Smedt, K., Meurer, P., and Dyvik, H. (2012). An open infrastructure for advanced treebanking. In *META-RESEARCH Workshop on Advanced Treebanking at LREC2012*, pages 22–29. Hajič, Jan.

Skjærholt, A. (2014). A chance-corrected measure of inter-annotator agreement for syntax. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 934–944, Baltimore, Maryland, June. Association for Computational Linguistics.

Solberg, P. E., Skjærholt, A., Øvrelid, L., Hagen, K., and Johannessen, J. B. (2014). The Norwegian Dependency Treebank. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2014)*, pages 789–795, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Stymne, S., de Lhoneux, M., Smith, A., and Nivre, J. (2018). Parser training with heterogeneous treebanks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia.

van der Goot, R. and de Lhoneux, M. (2021). Parsing with pretrained language models, multiple datasets, and dataset embeddings. In *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*, pages 96–104, Sofia, Bulgaria, December. Association for Computational Linguistics.

Velldal, E., Øvrelid, L., and Hohle, P. (2017). Joint ud parsing of norwegian bokmål and nynorsk. In *Proceedings of the 58th Conference on Simulation and Modelling*, pages 1–10. Linköping University Electronic Press.