

The Norwegian Dialect Corpus Treebank

Andre Kåsen¹ Kristin Hagen² Anders Nøklestad² Joel Priestley²
Per Erik Solberg¹ Dag Trygve Truslew Haug²

¹National Library of Norway ²Text Laboratory, Department of Linguistics and Scandinavian Studies, University of Oslo

Introduction

- The Norwegian Dialect Corpus (NDC) Treebank is a treebank of spoken Norwegian dialects from The Nordic Dialect Corpus (Johannessen et al. (2009) transcribed in the Bokmål variety of Norwegian.
- The NDC Treebank consists of 4587 speech segments, overall 66009 tokens, from 17 different Norwegian dialects from south, west, east and north of Norway, see Figure 1.
- The recordings in the corpus were made between 2006 and 2012 and comprise both interviews and more informal conversations between pairs of speakers.
- 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
 - The LIA Treebank of Spoken Norwegian Dialects (Øvrelid et al. 2018) with transcriptions in Nynorsk.

Annotation

- The annotation in the treebank follows the LIA Treebank, which extends the annotation scheme of NDT with a treatment of spoken-language phenomena.
- The treebank was preprocessed with an ad hoc pipeline for lemmatization, morphological features, part of speech and dependency syntax.
- Two linguistically trained annotators corrected the output of the morphosyntactic preprocessing using the [ConlluEditor](#) (Heinecke 2019).
- The annotation scheme aims at being as linguistically accurate as possible, following the Norwegian Reference Grammar (Faarlund et al. 1997).

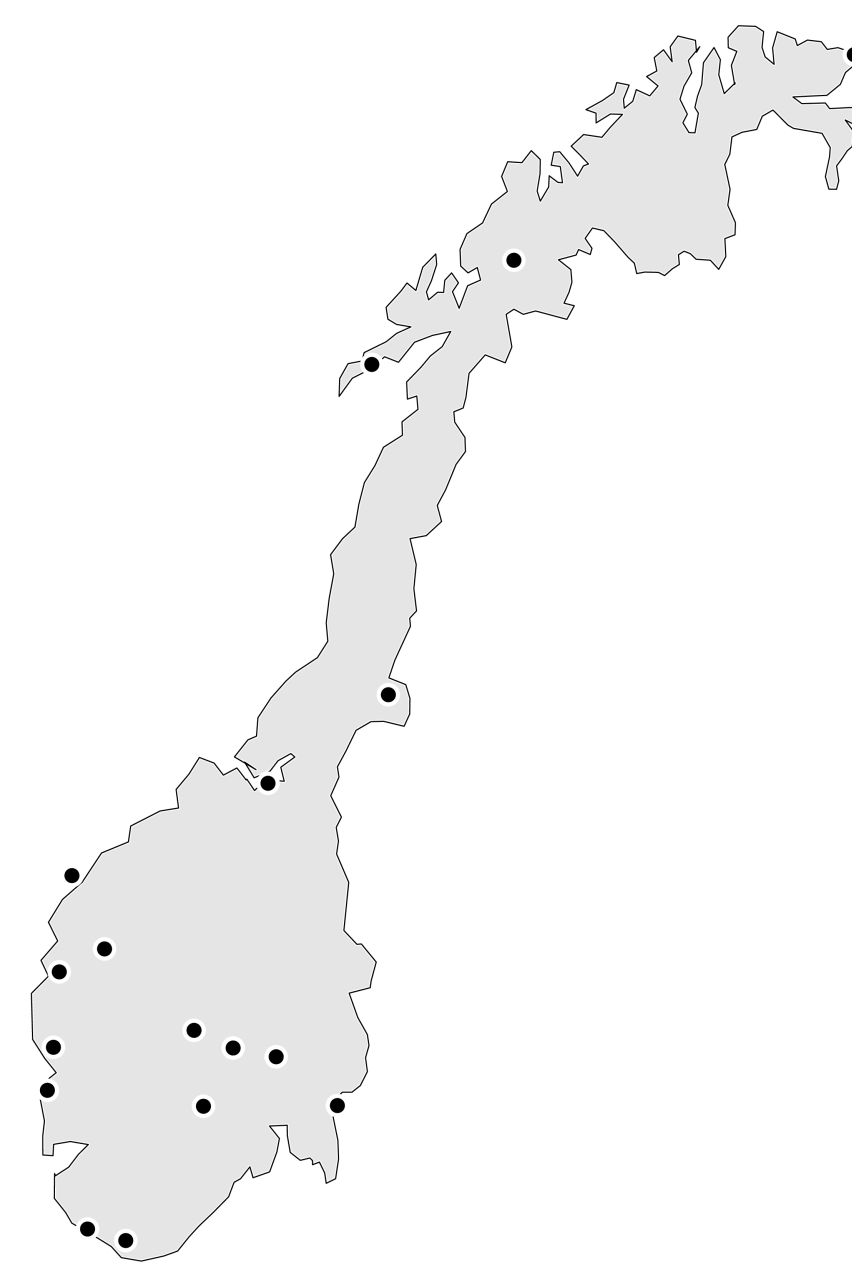
Project homepage

https://github.com/textlab/spoken_norwegian_resources

Experiments

- The manually corrected treebank was split with the UD guidelines for dataset release in mind (*If you have between 30K and 100K words, take 10K as test data, 10K as dev data and the rest as training data.*), and weighted for dialect.
- Two kinds of experiments were conducted: 1) in the style of Szymne et al. 2018 with the different treebanks for Norwegian, and 2) a cross validation inspired evaluation where every dialect in NDC served as a test set.
- Parser models were trained with UUParser (de Lhoneux et al. 2017).

Dialect distribution and results



Treebanks	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 set.

Figure 1: The transcriptions in the NDC Treebank are chosen from the same areas as the transcriptions in the LIA Treebank.

Accessibility

- The treebank is made available for search in Glossa, a web-based linguistic search interface with the ability to restrict the searches based on informant metadata.

Example of challenging segmentation and the SLETT relation

