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).

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 Stymne 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

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

<table>
<thead>
<tr>
<th>Treebanks</th>
<th>LAS</th>
<th>UAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDT</td>
<td>49.98</td>
<td>60.07</td>
</tr>
<tr>
<td>NDC</td>
<td>76.18</td>
<td>83.41</td>
</tr>
<tr>
<td>NDT + NDC</td>
<td>77.87</td>
<td>84.25</td>
</tr>
<tr>
<td>NDT + LIA + NDC</td>
<td>78.52</td>
<td>85.04</td>
</tr>
<tr>
<td>NDT + LIA + NDC</td>
<td>78.61</td>
<td>84.84</td>
</tr>
</tbody>
</table>

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

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