Building gold-standard treebanks for Norwegian

Per Erik Solberg, National Library of Norway, per.solberg@nb.no

THE TREEBANKS

- Språkbanken at the National Library of Norway is currently developing gold-standard Dependency Grammar treebanks for Norwegian Bokmål and Nynorsk
- Manually annotated for morphological features, syntactic functions and dependency relations
- Beta versions available at Språkbanken’s web page

Size of current beta version (0.4):
204 000 tokens for Bokmål and 156 000 tokens for Nynorsk

Distribution of texts:
- Newspaper text 79 %
- Government reports 9 %
- Parliament debates 6 %
- Blog posts 6 %

MORPHOLOGY AND LEMMATIZATION

- Lemmatization, POS-tags and morphological features mostly follow the Oslo Bergen Tagger (OBT)

<table>
<thead>
<tr>
<th>Form</th>
<th>Lemma</th>
<th>POS</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Se</td>
<td>se</td>
<td>verb</td>
<td>imp</td>
</tr>
<tr>
<td>innslag</td>
<td>innslag</td>
<td>subst</td>
<td>appell</td>
</tr>
<tr>
<td>her</td>
<td>her</td>
<td>prep</td>
<td>–</td>
</tr>
<tr>
<td>.</td>
<td>$.</td>
<td>clb</td>
<td>&lt;punkt&gt;</td>
</tr>
</tbody>
</table>

- Non-standard forms are marked with the additional tag unorm

<table>
<thead>
<tr>
<th>Form</th>
<th>Lemma</th>
<th>POS</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>kallet</td>
<td>kalle</td>
<td>verb</td>
<td>perf-part</td>
</tr>
</tbody>
</table>

SYNTACTIC ANNOTATION

- Syntactic annotation guidelines have been developed specifically for this project

(1)

Den kinesiske lederen har tydeligvis innsett alvoret.

The Chinese administration the has evidently realized seriousness+the

‘The Chinese administration has evidently realized the seriousness.’

(2)

De kritiske røstene hevdet at foreldrene overstimulerte barna.

The critical voices claimed that parents+the overstimulated children+the

‘The critical voices claimed that the parents overstimulated the children.’

USE

- Testing and training of NLP tools
  - We have obtained a labeled attachment score of 85.77 % from the MaltParser, a data-driven dependency parser, trained on the Bokmål treebank

Linguistic research

- A useful tool for research within linguistic disciplines such as morphology, lexicography and syntax
  - It is possible to make queries for very specific syntactic constructions, e.g. fronted non-subjects in complement clauses
  - You can retrieve all lemmas which occur with the same syntactic function, e.g. indirect object

Contact

Inquiries regarding the treebanks can be sent to sprakbanken@nb.no

Contact

Inquiries regarding the treebanks can be sent to sprakbanken@nb.no