Nasjonalbiblioteket

Norwegian Dependency Treebank version 2.0.

The Norwegian Language Bank

This is version 2.0 of the Norwegian Dependency Treebank (NDT), developed by the National Library of Norway in 2011-2014. The original version can be found here: https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-10/

The treebank is also converted to Universal Dependencies. The latest releases of the UD version can be found here:
https://universaldependencies.org/treebanks/no_bokmaal/index.html (Norwegian Bokmål)
https://universaldependencies.org/treebanks/no_nynorsk/index.html (Norwegian Nynorsk)

In version 2.0 of the NDT, the grammatical annotations remain the same as in the previous NDT version, but information useful metadata from the UD version of the treebank has been added, and an effort has been made to make the word tokenization and sentence segmentation of the UD and NDT versions exactly parallel. The treebank has also been split into a test, train and evaluation set following the UD splits.

These are the modifications in v. 2.0:

- The treebank files are converted from the CONLL-X format to the CONLL-U format.
- A few modifications have been made to the word tokenization and the sentence segmentation to ensure that the tokenization and segmentation corresponds exactly to the UD version of the treebank. It is important to note that these modifications make the tokenization and segmentation more consistent compared to the previous version.
- Before each sentence, a *sentence_id* and a *text* comment are added. Both are taken from the UD version of the treebank. The *text* is the text string of the sentence as it appeared in the original text.
- Before sentences marking the start of a new paragraph in the original text, there is a *newpar* comment. This information is also taken from the UD version of the treebank.
- In the previous release, a pipe symbol, “|”, was added at the end of sentences which didn’t end with punctuation. These pipes have, for the most part, been removed. However, there are a few remaining instances in the UD version, and in those cases, the pipe remains also in this version so that the UD and the NDT versions remain parallel.
- The Bokmål and Nynorsk treebanks have been split in test, train and evaluation sets. These splits follow the splits in the UD version.

If you have questions or comments, please feel free to contact us at sprakbanken@nb.no.