Beta version of Språkbanken's gold standard corpus for Norwegian Bokmål and Nynorsk

The corpora are under development at the National Library of Norway/Språkbanken in cooperation with the Text Laboratory at the University of Oslo. This version is from August 6, 2012.

Contents

20120806_nob_gullkorpus.conll	Gold Standard Corpus for Norwegian Bokmål
20120806_nno_gullkorpus.conll	Gold Standard Corpus for Norwegian Nynorsk

On the file format

The corpora are in the conll format. For more on this format, we refer to http://ufal.mff.cuni.cz/conll2009-st/task-description.html.

Brief on the content

The corpora consist of text which is manually annotated with morphological features, syntactic functions and hierarchical structure. The morphological annotation in large part follows the Oslo-Bergen tagger (http://tekstlab.uio.no/obt-ny/english/index.html). The formalism used for the syntactic annotation is dependency grammar. With a few exceptions, the syntactic analysis follows *Norsk referensegrammatikk* 'Norwegian Reference Grammar' (Jan Terje Faarlund, Svein Lie and Kjell Ivar Vannebo, Universitetsforlaget 1997). A detailed annotation manual will be published at a later stage.

Text selection

Thus far, the corpora comprise of newspaper text from the late 90ies to the present day and parliamentary proceedings from the same time period.

Norwegian Bokmål	Norwegian Nynorsk
Aftenposten	Dag og Tid
Dagbladet	Vest-Telemark Blad
Klassekampen	Klassekampen
Parliamentary proceedings	

A detailed overview of the text contents will be added later.

For questions, comments and bug reports, please contact Språkbanken at sprakbanken@nb.no.