Corpus specifications

The ingredients

DK-CLARIN WP 2.1 Technical Report.
Draft version of March 5, 2014[1]

Deliverables concerned

D18 Final version of corpus Final version of POS-tagged corpus of 45 million words available for the DK-CLARIN repository and accessible through a web-based (or other) concordance tool. **Outcome:** Resource with documentation.

Outline

This technical report describes the composition of the corpus, the text material included, and how the corpus can be accessed.

Document history

The most recent version may have important content modifications. You can download it from:

- http://korpus.dsl.dk/clarin/corpus-doc/corpus-specs.pdf.

[1] A more recent version may be available at:
http://korpus.dsl.dk/clarin/corpus-doc/corpus-specs.pdf

1 Corpus composition

The following table shows from which sources the text material included in the DK-CLARIN corpus were drawn.

Type	Source	Period	Capture	Text Items	Tokens	Remarks
bg	Bentes blog	2008–2011	dsn.dk			
bg	Blogbogstaver	2005–2011	dsn.dk			
bg	Blogsbjerg	2007–2011	dsn.dk			

2 Text material

2.1 Wikipedia

- Headlines are left out in wikipedia articles which constitutes a severe problem as certain parts of these articles not relevant in a corpus context not can be filtered out, especially the references and links.
- List and tables articles have not been removed. Table articles contain lots of | chars that were erroneously classified as words but have been reclassified into punctuation characters by applying textjuggler.TextCleaner.
- Space characters in <title> elements are erroneously replaced by underscores. Some titles end with a hex number.

3 Corpus access