Building a DDC Annotated Corpus from OAI Metadata

Mathias Lösch[1], Ulli Waltinger[2], Wolfram Horstmann[3], and Alexander Mehler[2]

[1] Bielefeld University Library Mathias.Loesch@uni-bielefeld.de
[2] Text Technology / Applied Computational Linguistics, Bielefeld University {Ulli_Marc.Waltinger,Alexander.Mehler}@uni-bielefeld.de
[3] Bielefeld University Wolfram.Horstmann@uni-bielefeld.de

Abstract. A frequently overlooked benefit of open access publications is that they are an easy accessible and cost-effective data source for research disciplines like text mining, natural language processing or computational linguistics. In those fields, linguistic data is usually managed in the form of *corpora*, i.e. machine readable bodies of texts that represent a particular variety of language.

The contribution presents a bilingual (English and German) corpus consisting of OAI records that were originally harvested for the *Bielefeld Academic Search Engine* (BASE) and their respective full texts. A particular added value is that every record is annotated with at least one Dewey Decimal Classification (DDC) code. This was achieved by the use of data analysis and mapping tables from subject-specific classification schemes to the DDC and thus, all texts in the corpus are accessible via their subjects.

The presented work is part of the project *Automatic Enhancement of OAI Metadata*, which is a cooperation between Bielefeld University Library, the Department of Computational Linguistics at Bielefeld University and the Department of Computer Science at the University of Leipzig, and which is funded by the *German Research Foundation* (DFG).

Keywords

Dewey Decimal Classification, OAI metadata, corpus construction