DeutschDiachronDigital - A Diachronic Corpus of German

Anke Lüdeling, Humboldt-University Berlin

In my talk I present the concept for the design and the architecture of a diachronic corpus of German, as developed in the project initiative *DeutschDiachronDigital* (henceforth DDD).

There are many digitized historical German texts from all periods (Old High German to Modern German). It is, however, difficult to carry out diachronic research because

- there are differences in digitzation source (original or edition)
- there are differences in digitization quality
- the texts are stored in different (and, sometimes, incompatible) formats
- many texts are not publicly available
- very often, header data is missing or not encoded according to standards
- very often, there is no linguistic annotation
- often, documentation is not sufficient
- corpus design differs for the different linguistic periods.

Because of this situation, scholars from 15 universities and research institutions in Germany together with several international partners decided to form the DDD initiative to build a homogeneous diachronic corpus of German for all textual sciences (linguistics, literature, philology, history etc.).[1]

The requirements for such a corpus are

(1) homogeneity with respect to design, digitization, common meta-information and common annotation layers
(2) flexibility with respect to additional texts, meta-information, and annotation layers.

Homogeneity can be achieved through thorough standardization of digitization guidelines, tagsets, annotation guidelines, and procedures. DDD complies with all common standards (TEI, IMDI, OLAC, Unicode (Menota)), but many decisions still have to be made. Corpus design is based on the parameters time (50 year slices), dialect/regional variety, and text type. **Flexibility** can be achieved in a multi-layer standoff corpus architecture, in an extended ODAG format (Carletta et al. 2003, Dipper et al. 2004, Faulstich, Leser & Lüdeling 2006). Texts and annotations are stored in XML. The smallest annotation unit is the character. All annotation layers are stored in separate layers which point to spans in the text. The corpus and all annotation layers are stored in a relational database (see abstract by Leser & Karosseit in these proceedings). It is possible to store several aligned 'views' or representations of the same text (for example, a diplomatic version of the text and a normalized version) which each have their own annotation layers (Lüdeling, Poschenrieder & Faulstich 2006).

Carletta, Jean; Kilgour, Jonathan; O'Donnell, Timothy; Evert, Stefan & Voormann, Holger (2003) The NITE object model library for handling structured linguistic annotation on multimodal data sets. In *Proceedings of the EACL Workshop on Language Technology and the Semantic Web* (3rdWorkshop on NLP and XML, NLPXML-2003).

Dipper, Stefanie; Faulstich, Lukas; Leser, Ulf & Lüdeling, Anke (2004) Challenges in Modelling a Richly Annotated Diachronic Corpus of German. In: *Workshop on XML-based richly annotated corpora*, Lisbon, Portugal

Faulstich, Lukas C; Leser, Ulf & Lüdeling, Anke. *Storing and Querying Historical Texts in a Database*. Technical Report 176 des Instituts für Informatik der Humboldt-Universität zu Berlin, January 2005.

Lüdeling, Anke; Poschenrieder, Thorwald & Faulstich, Lukas C. (2005) DeutschDiachronDigital - Ein diachrones Korpus des Deutschen. In: Jahrbuch für Computerphilologie 2004, 119-136. Available online at http://computerphilologie.uni-muenchen.de/ejournal.html

[1] A first funding application was rejected by the Deutsche Forschungsgemeinschaft. The initiative will continue its effort to receive funding.