GerManC: An Annotated, Spatialised, Multi-Genre Corpus of Early Modern German

Martin Durrell,[1] Paul Bennett
and Astrid Ensslin

Abstract

The authors are currently engaged in a project which aims to compile a representative annotated corpus of German for the period 1650–1800 ('GerManC'). This corpus is 'spatialised' temporally and topographically, i.e. subdivided into three 50-year spans (1650–1700, 1701–1750, 1751–1800) and the five major dialectal regions of the German Empire. In a one-year pilot study it proved possible to identify the challenges posed by the particular lexical, morphological, syntactical and graphemic peculiarities characteristic of that particular stage of written German (Durrell et al., in preparation). It is now proposed to apply these findings to a more extensive 800,000 word historical corpus comprising eight written genres (personal letters, drama, sermons; newspapers, academic, medical, and legal texts as well as prose fiction); the compilation of this is just starting. In line with comparable international corpus projects, multi-layer stand-off annotation will be applied, and the feasibility of integrating this kind of architecture with a relational database approach will be investigated. Of further interest will be the possibility of semi-automatising TEI annotation and of applying innovative spelling variant retrieval methods to the large amount of allographic variation found in pre-standard German (cf. Ernst-Gerlach and Fuhr, 2006; Pilz et al., 2006).

References

Durrell, Martin, Paul Bennett and Astrid Ensslin (in preparation) 'GerManC – Towards a Methodology for Constructing and Annotating Historical Corpora', in *Proceedings of the 'Digital Historical Corpora - Architecture, Annotation, Retrieval' Conference*, *Dagstuhl, Dec. 2006*. PPT available at http://kathrin.dagstuhl.de/06491/Materials2/ (13 December 2006).

Ernst-Gerlach, Andrea and Norbert Fuhr (2006) 'Generating Search Term Variants for Text Collections with Historic Spellings', in *Proceedings of the 28th European Conference on Information Retrieval Research (ECIR 2006)*. Available at: www.is.informatik.uni-duisburg.de/bib/pdf/ir/Ernst_Fuhr:06.pdf (13 December 2006).

Pilz, Thomas, Wolfram Luther, Ulrich Ammon and Norbert Fuhr (2006) 'Rule-based Search in Text Databases with Nonstandard Orthography'. *Literary and Linguistic Computing*, 21(2), 179–86.

[1] University of Manchester
e-mail: martin.durrell@manchester.ac.uk