Abstract for 'Digital Historical Corpora - Architecture, Annotation, and Retrieval' Conference, 03-08 December 2006, Dagstuhl (D)

GerManC- Towards a Methodology for Constructing and Annotating Historical Corpora

Astrid Ensslin, Martin Durrell, Paul Bennett
University of Manchester (UK)

Our paper focuses on the one hand on the challenges posed by the structural variability, flexibility and ambiguity found in historical corpora and evaluates methods of dealing with them on the other.

We are currently engaged in a project which aims to compile a representative corpus of German for the period 1650-1800. Looking at exemplary data from the first stage of this project (1650-1700), which consists of newspaper texts from this period, we first aim from the perspective of corpus linguistics to identify the problems associated with the morphological, syntactical and graphemic peculiarities that are characteristic of that particular stage. Specific phenomena which significantly complicate automatic tagging, lemmatisation and parsing include, for instance, 'abperlende' (Admoni 1980; Demske-Neumann 1990), i.e. complex and often asyndetic syntax; non-syntactic, prosodic, virgulated punctuation (Demske et al. 2004; cf. Stolt 1990), inflectional variability (e.g. Admoni 1990; Besch & Wegera 1987), as well as partly unsystematic and almost experimental allomorphic and allographic (Kettmann, 1992) diversity.

Secondly, we outline a methodology which is intended to facilitate the construction and annotation of such corpora which antedate linguistic standardisation. This is informed by 'conventional' and innovative tagging techniques and tools, which are evaluated in terms of utility and accuracy. Finally, we attempt to evaluate the degree to which annotation tools for specialist corpora of this kind can be developed which will substitute for manual or semi-automated annotation.

References:

Admoni, Wladimir (1980) *Zur Ausbildung der Norm der deutschen Literatursprache im Bereich des neuhochdeutschen Satzgefüges (1470 - 1730)*. Berlin: Akademie-Verlag.

Besch, Werner & Wegera, Klaus Peter (1987) *Frühneuhochdeutsch. Zum Stand der sprachwissenschaftlichen Forschung*, *special edition of Zeitschrift für deutsche Philologie*, 106.

Demske, Ulrike, Frank, Nicola, Laufer, Stefanie & Stiemer, Hendrik (2004) 'Syntactic Interpretation of an Early New High German Corpus' in S. Kübler et al. (eds.) *Proceedings of the Third Workshop on Treebanks and Linguistics Theories (TLT 2004)*, pp. 175-182. Tübingen.

Demske-Neumann, Ulrike (1990) 'Charakteristische Strukturen von Satzgefügen in den Zeitungen des 17. Jh.' in A. Betten (ed.) *Neuere Forschungen zur historischen Syntax*

des Deutschen. Referate der Internat. Fachkonferenz Eichstätt 1989, pp. 239-252. Tübingen.

Kettmann, Gerhard (1992) 'Zum Graphemgebrauch in der dt. Literatursprache. Variantenbestand und Variantenanwendung (1570-1730)' in J. Schildt (ed.) *Soziolinguistische Aspekte des Sprachwandels in der dt. Literatursprache 1570-1730*, pp. 15-118. Berlin: Akademie-Verlag.

Stolt, Birgit (1990) 'Redeglieder, Informationseinheiten: *Cola* und *commata* in Luthers Syntax' in A. Betten (ed.) *Neuere Forschungen zur historischen Syntax des Deutschen. Referate der Internat. Fachkonferenz Eichstätt 1989*, pp. 377-390. Tübingen.