Abs-173

Silke Scheible, Paul Bennett, Martin Durrell, and Richard J. Whitt (all University of Manchester)

The GerManC project: Creating an annotated historical corpus of German

This paper describes the results of an AHRC/ESRC-funded project on the creation of a corpus of Early Modern German scheduled for completion in September 2011. GerManC aims to be a representative corpus of written German between 1650 and 1800, and includes a total of eight different genres (representing both print-oriented and orally-oriented registers). It is further subdivided into five dialect regions and three 50-year time periods. The corpus consists of sample texts of 2,000 words, equally distributed across the 'genre', 'period', and 'region' categories, yielding a total of nearly a million words. The structure of the corpus resembles the design of the ARCHER corpus for English. Thus, GerManC is not only of interest for historical German linguists, but it also promises to be an important resource for comparative studies of the development of the two languages.

Our paper reports on the challenges encountered in compiling the corpus, which involves identifying suitable texts and digitising texts printed in Fraktur ('black letter', 'Gothic'). It further provides a detailed account of the annotation of the corpus in terms of structural mark-up (TEI) and linguistic mark-up (sentence boundaries, normalised word forms, lemmas, and POS-tags). The presentation will focus on the following major goals of the annotation process: a.) Creating an automatic linguistic annotation pipeline which is especially suited to historical German in this period, and b.) identifying strategies for a speedy manual correction of the errors produced by the automatic tools. Both points are of vital importance for maximising the overall accuracy of the annotations in the corpus.

To address a.), a novel tokenizer and sentence boundary detector have been created which can deal with multi-genre input such as found in our corpus. Furthermore, we will describe experiments carried out on a manually-annotated subcorpus of GerManC (ca. 50,000 tokens), whose aim is to maximise the performance of state-of-the-art POS-taggers such as the TreeTagger and the TnT Tagger for German. We will report the results of running the taggers on 'raw' vs. 'normalised' data, and compare these findings to the performance of the taggers when re-trained on our gold-standard subcorpus. The results of these experiments are utilised for creating a historical text processing pipeline optimised for historical input, which minimises the amount of manual correction necessary for a gold standard annotation of the corpus. Finally, addressing point b.), we will introduce a novel web-based annotation platform which is currently being developed as part of the project. Its purpose is to facilitate fast and easy manual correction of token-based annotations such as lemmas and POS tags. The platform will be freely available, and we plan to give a short demonstration of its functionalities at the end of our presentation.