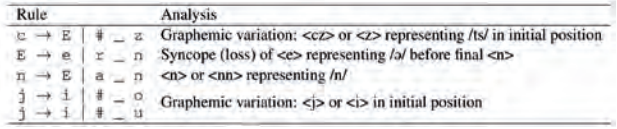
Towards a broad-coverage graphematic analysis of large historical corpora

Mittwoch,
04.03.2010
15:15–15:45
VMP5 2098/2194

Stefanie Dipper, Ilka Lemke, Sandra Waldenberger
University of Bochum
dipper@linguistics.rub.de, ilka.lemke@rub.de, sandra.waldenberger@rub.de

AG 5

This contribution presents a set of methods we develop to explore graphemic and graphematic variation in large historical corpora of German, starting with the *Referenzkorpus Mittelhochdeutsch*. We apply methods from computational linguistics to pave the way for a broad-coverage graphematic analysis. In essence, we use the normalization level provided by the annotations in said corpus, first, to automatically identify 'equivalent' word forms in two texts from different language areas or time periods (e.g. *czeit* and *zeit* 'time'), and, secondly, to derive replacement rules and mappings from these word forms (cf. Dipper/Waldenberger 2017). Some example rules derived this way from the *Anselm Corpus* (cf. Dipper & Schultz-Balluff 2013) are shown in the following figure from Dipper/Waldenberger (2017: 40). The rule notation uses 'E' to represent the empty string and '#' for the word boundary. For instance, the first rule 'c → E | # _ z' effectively deletes a word-initial 'c' in front of a 'z'.



<table>
<tr><th>Rule</th><th>Analysis</th></tr>
<tr><td>c → E | # _ z</td><td>Graphemic variation: &lt;cz&gt; or &lt;z&gt; representing /ts/ in initial position</td></tr>
<tr><td>E → e | r _ n</td><td>Syncope (loss) of &lt;e&gt; representing /ə/ before final &lt;n&gt;</td></tr>
<tr><td>n → E | a _ n</td><td>&lt;n&gt; or &lt;nn&gt; representing /n/</td></tr>
<tr><td>j → i | # _ o</td><td rowspan="2">Graphemic variation: &lt;j&gt; or &lt;i&gt; in initial position</td></tr>
<tr><td>j → i | # _ u</td></tr>
</table>

The replacement rules derived by our scripts are then analysed using expertise in historical German linguistics (see column 'Analysis' in the figure). This approach offers decisive advantages compared to existing approaches which either rely on a predefined set of characteristics (as summarized in Elmentaler 2018: 328–336) or have had to restrict themselves to a relatively small and limited corpus (cf. e.g. Moser 1977; Glaser 1985; Rieke 1998; Elmentaler 2003). Basically, we offer an approach that answers Elmentaler's (2018: 335) call for a (semi-)automatic analysis of graphemic variables.

References: Anselm-Corpus: https://www.linguistics.rub.de/comphist/projects/anselm/; Dipper, S. & S. Schultz-Balluff (2013). The Anselm Corpus: Methods and Perspectives of a Parallel Aligned Corpus. In Proceedings of the NODALIDA Workshop on Computational Historical Linguistics (NEALT Proceedings Series; 18). Oslo, Norway, 27–42. Dipper, S. & S. Waldenberger (2017). Investigating Diatopic Variation in a Historical Corpus. In Proceedings of the EACL-Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), 36–45. Valencia, Spain: Association for Computational Linguistics. Elmentaler, M. (2003). Struktur und Wandel vormoderner Schreibsprachen. Berlin, New York: De Gruyter. Elmentaler, M. (2018). Historische Graphematik des Deutschen. Eine Einführung. Tübingen: Narr. Glaser, E. (1985). Graphische Studien zum Schreibsprachwandel vom 13. bis 16. Jahrhundert. Vergleich verschiedener Handschriften des Augsburger Stadtbuchs. Heidelberg: Winter. Moser, H. (1977). Die Kanzlei Kaiser Maximilians I. Graphematik eines Schreibusus. Innsbruck. Referenzkorpus Mittelhochdeutsch [reference corpus Middle High German]: T. Klein, K.-P. Wegera, S. Dipper, & C. Wich-Reif (2016); Referenzkorpus Mittelhochdeutsch (1050–1350), Version 1.0. https://www.linguistics.ruhr-uni-bochum.de/rem/. ISLRN 332-536-136-099-5. Rieke, U. (1998). Studien zur Herausbildung der neuhochdeutschen Orthographie. Die Markierung der Vokalquantitäten in deutschsprachigen Bibeldrucken des 16.–18. Jahrhunderts. Heidelberg: Winter.