Towards a broad-coverage graphematic analysis of large historical corpora

Stefanie Dipper, Ilka Lemke, Sandra Waldenberger

University of Bochum
dipper@linguistics.rub.de, ilka.lemke@rub.de, sandra.waldenberger@rub.de

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’.

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.