Quantifying graphematic variation via large text corpora

Hanna Lüschow

Carl von Ossietzky Universität Oldenburg

hanna.lueschow@uol.de

Mittwoch,
04.03.2020
16:30–17:00
VMP5 2098/2194

This talk discusses the use of some basic computer science concepts for expanding the possibilities of (manual) diachronic graphematic text corpus analysis.

With these it can be shown that graphematic variation decreases constantly in printed German texts from 1600 to 1900. While the variability is continuously lesser on a text-internal level, it decreases faster for the whole known writing system of individual decades.

But which changes took place exactly? Which types of variation went away more quickly, which ones persisted? How do we deal with large amounts of data which cannot be processed manually? Which aspects are of special importance/go missing when working with a large textual base?

The use of a measure called entropy (Shannon 1948) quantifies the variability of the spellings of a given wordform, lemma, text or subcorpus, with few restrictions but also less details in the results. The difference between two spellings can be measured via Damerau-Levenshtein distance (Damerau 1964, Levenshtein 1966). To a certain degree, automated data handling can also determine the exact differences at hand. Afterwards, these differences can be counted and ranked.

As data source the German Text Archive of the Berlin-Brandenburg Academy of Sciences and Humanities is used. It offers for example orthographic normalization (which proved to be very useful), preprocessing of parts of speech and lemmatization.

These methodological findings could subsequently be used for improving research methods in other graphematic fields of interest, for cross-linguistic as well as for non-diachronic data. For a comparison to other languages, we ‘only’ need large amounts of data with similar preprocessing; the methodological approaches should remain rather consistent. The same holds for analyzing computer-mediated communication (or anything else with at least a little variation).

Natural Language Processing (NLP)-tools for analyses below word-level aren’t really widespread. At a later point, more advanced techniques from the realm of natural language processing and/or machine learning could be used or even newly developed. Therefore, this approach also strongly advocates for interdisciplinarity.

References: Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. Communications of the ACM 7(3), 171–176. Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10(8), 707–710. Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal 27(3), 379–423.