Measuring Syntactic Change: Underuse and Overuse Statistics in a Multi-Layer Historical Corpus of German

Abstract

This paper examines methodological issues in the study of diachronic language change using empirical data, with examples taken from different periods of High German language. Using the variationist approach, we assume that the grammar of a language in different, closely related stages does not develop primarily through abrupt qualitative changes, but rather through quantitative changes in the functions of competing forms. These forms can be regarded as different variants of a variable, which corresponds to a linguistic function, or put more loosely, they code 'different ways of saying the same thing'. The study of such variables requires first and foremost a formalization of their variants and an answer to the basic question 'what is the same thing'. At the same time, variables found in empirical corpus data must be coded in a way that allows comparability between language stages despite the differences between them. Finally, we wish to find out which variables are chiefly responsible for the differences between language stages.

To deal with the first challenge, we take an approach based on the use of a multi-layer corpus architecture, using the ANNIS2 corpus search and visualization system. Since the definition of variables depends on the coding of their variants we compare and contrast multiple annotation levels grouping together different competing forms at the morphological, syntactic and lexical levels. Richly annotated multi-layer corpora allow us to represent the resulting conflicting structures while adding new forms of annotation as they become necessary. It then becomes possible to investigate the interrelation between different annotation levels as well.

Once we establish synchronic annotations for different variables, we must ensure that the annotation scheme remains comparable across language periods. This is a prerequisite for enabling quantitative comparisons diachronically. The difficulty lies in the fact that categories change across time: Old High German, for instance, is said not to have had definite articles, but the forerunners of these elements are found in the form of demonstratives. We attempt to solve such tensions through multiple annotation layers, e.g. syntactic (constituents and grammatical function) and morphological (part of speech and inflection), where an aspect of comparability is retained in the one layer, while a difference is expressed on the other. We also discuss the use of synchronic lemmatization and diachronic hyper-lemmatization (giving common lemmas to historically related items) to allow better lexical comparisons across time. Throughout this discussion, maximal comparability of annotation schemes is the key concern.

With comparable corpora encoding relevant variables at hand, we proceed to diagnose the most prominent features of different periods as well as changes which proceed gradually through time. We do this using underuse / overuse descriptive statistics, searching for and visualizing the space of categories whose frequency deviates significantly between language stages. These diagnostics are then used as a point of access for more detailed qualitative analyses.