Stefan Evert

A multivariate approach to linguistic variation and distribution

Most linguistic features exhibit considerable variability in their frequency distribution across different time periods, language varieties, regions, speakers, text types, topic domains, etc. Such language variation is of great interest to corpus linguistics, sociology, dialectology, historical linguistics, and many other fields of research. However, it can also be a confounding factor for studies that focus on a specific contrast, obscuring or distorting the relevant differences, and needs to be accounted for in a statistical analysis.

In many cases, several related or unrelated linguistic features follow a similar pattern of variation. Such correlations can be analysed with multivariate statistical techniques – such as factor analysis, principal component analysis and correspondence analysis – in a completely unsupervised manner, revealing insights about the dimensions of linguistic variation as well as the distribution of individual features. A well-known example of this inductive approach is Doug Biber's multidimensional register analysis. Computational linguists use similar techniques in order to identify topic domains and obtain a distributional representation of word meaning.

In my talk, I will present case studies of multivariate analyses, highlighting both their common core and individual differences between the approaches. I will then focus on several important methodological problems, in particular (i) the possibility of introducing researcher bias through the choice of text samples and features and (ii) the difficulty of assigning a meaningful interpretation to the identified dimensions of variation. Recent findings suggest that a minimal amount of supervision – e.g. in the form of language-external categories – can help to guide the multivariate analysis towards interpretable dimensions representing specific aspects of language variation.