There is a significant gap between the high level of abstraction of linguistic models, such as SFL (Fawcett 2006, Halliday and Matthiessen 2014), and data provided through shallow analysis and annotation of electronic corpora (Alves et al 2010, Steiner 2012, Hansen-Schirra, Neumann and Steiner 2012, Kunz et al. 2017). This is one of the reasons why work in my own group has used SFL for generating hypotheses and for interpreting theory-neutral data, but rarely as annotations in the data directly. Direct SFL-annotations are a) costly, b) inconsistent between annotators, and c) make the data theory-dependent.

In a range of different studies using data-mining techniques (Taboada et al 2011, Degaetano-Ortlieb et al 2014), attempts have been made to use ideas from SFL as an underlying linguistic model to some extent. The latter use 3 levels of analysis (shallow features, a limited set of features from register theory, and finally a combination of these). The combination shows improved results for text classification compared to pure “bag-of-words”-type approaches, at least in situations in which linguistic differences of registers are marked. Direct SFL-annotations in data are possible, at least for well-operationalized register features. There is a question of how SFL-specific (some of) these features really are.

In order to further narrow the gap between SFL-theorizing and data, improved strategies have to be developed of formulating empirically-testable hypotheses. Two examples of studies of cohesive chains will be outlined to approach this goal (Kunz et al 2016, Laphinova-Koltunski et al 2016). Yet once more the question is how SFL-specific (some of) these hypotheses and their operationalizations are.

Particular challenges for SFL as an underlying model result from the fact that SFL annotations constitute highly interpreted data. They are difficult in terms of inter-coder consistency and expensive to create in sufficient quality. The resulting problems for outside evaluation and repeatability of studies, one important way of enhancing quality of empirical research, need to be addressed.

References:


