Measuring variation in English and beyond

This talk introduces three recent quantitative approaches to measuring variation in corpora:

(1) **Multivariate logistic regression analysis** is used to assess the constraints governing the English genitive variation (e.g. *Tom's car* vs. *the car of Tom*). The focus of analysis is put on the rarely explored factor "rhythm" in Early Modern English drawing on data from ARCHER (A Representative Corpus of Historical English Registers). According to the Principle of Rhythmic Alternation (Schlüter 2005) more rhythmically optimal genitive constructions should be preferred to the less rhythmically optimal variant. Yet, it turns out that rhythm is only a minor player in this dataset whose effect runs contrary to expectations as more rhythmic genitives are overall not preferred (Ehret et al. 2014).

(2) Lexical variation between written standard British and American English is measured in a bottom-up fashion by using **Semantic Vector Space Models** to generate a large dataset of lexical variables. The database is the Brown family of corpora which comprise two registers (informative vs. imaginative) for two time periods (1960's vs. 1990's) across two varieties (English vs. American English). It is demonstrated how lexical variation is systematically related to these three dimensions. Furthermore, many lexical variables are sensitive to one of these language-external dimensions (Ruette et al. 2016).

(3) The complexity of learner essays in ICLE (International Corpus of Learner English) is approximated via their **Kolmogorov complexity**. The basic idea is to use compression algorithms to assess language complexity by measuring the information content in text samples. Texts that can be compressed more efficiently are comparatively more complex. The current analysis focuses on German learner essays and shows how complexity generally increases with increasing levels of instruction in English (Ehret and Szmrecsanyi 2016).

To sum up, I will demonstrate how these methodologies can be used to explore different concepts of variation in English corpora. However, these methods need not be restricted to English datasets but can be applied to all kinds of variational studies in languages beyond English.

References

