Automatic register annotation for linguistic research?

Register information is among the types of annotation most sought for by many corpus linguists. Yet more often than not, corpora lack this kind of meta data. Recent experiments (e. g., Biber et al., 2015; Asheghi et al., 2016) show that identifying linguistically relevant register categories and operationalizing them such that texts can be categorized with satisfactory inter-rater reliabilility is an unsolved issue. Regardless of conceptual problems, register meta data can be manually added to corpora of moderate size, or it may sometimes even be implicit in the sampling scheme of a corpus. Manual annotation is not an option, however, for very large corpora, such as crawled web corpora. Obviously, the only viable alternative for enriching very large corpora with register meta data is automatic classification. Unfortunately, even very recent experimental results from automatic genre/register classification are rather unsatisfying for documents from unrestricted domains. For instance, Biber and Egbert (2015) report a mere 42.1% of correctly classified web documents, which is clearly insufficient for linguistic research. Moreover, most approaches to automatic genre/register classification rely on a (usually high) number of linguistic features extracted from the documents. This is also true for approaches that do not assume a pre-defined set of categories, but instead identify registers statistically in a bottom-up fashion (e. g. Biber, 1988). It is unclear how the resulting categories can be used in linguistic research without running into circularity (if, for instance, feature F is found to be particularly frequent in register R, when F was among the features used to establish register R in the first place).

In our talk, we explore the usefulness of automatically annotated register categories in modeling alternation phenomena in morpho-syntax. Our aim is to show that even if one is willing to accept potentially circular definitions of register categories derived from document-internal features, one might still end up with less optimal statistical models than could be obtained using the raw features. To this end, we extract a large number of features at the document level (such as relative frequencies of PoS tags, morphological markers, syntactic constructions, text length, type/token ratio, emoticons) and use them in a generalized linear model (GLM) to predict the occurrence of particular variants in alternation phenomena. We then use different methods to reduce the dimensionality of our dataset. First, we apply factor analysis to single out the most relevant dimensions of variation, in the spirit of Biber (1988) and subsequent work. We use the documents' loadings on each one of these factors as predictors in a second GLM, thus modelling the alternation phenomenon on a smaller set of predictors. Finally, we use hierarchical clustering (on the original features) in order to group the documents into a relatively small number of distinct classes. This last scenario corresponds to the case where documents are assigned to a single register category. In a third GLM, only these document classes are used as predictors.

We will apply this methodology in two case studies (on the basis of 160,000 documents, sampled from a web corpus and a corpus containing predominantly newspaper texts), focusing on well-known alternation phenomena in German:

1. dative/genitive case alternations after prepositions, such as *wegen* 'because of', *gemäß* 'according to', *einschließlich* 'including'
2. inflection of adjectives after pronominal adjectives, such as *mit manch leckerem Kuchen* vs. *mit manchem leckeren Kuchen* vs. *mit manchem leckerem Kuchen* 'with many a delicious cake'

In a third study, we turn to artificially generated data from a Monte Carlo simulation, which allows us to analyze the effects of subtle manipulations of population parameters. We compare the quality of the models thus obtained (no aggregation vs. partial aggregation vs. full aggregation) and discuss the implications of our findings for using automatically annotated high-level categories, such as register, in research on grammatical and morphological alternation phenomena.

Asheghi, N., Sharoff, S., and Markert, K. (2016). Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, 50:603–641.

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

Biber, D. and Egbert, J. (2015). Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Journal of Research Design and Statistics in Linguistics and Communication Science*, 2:3–36.

Biber, D., Egbert, J., and Davies, M. (2015). Exploring the composition of the searchable web: a corpus-based taxonomy of web registers. *Corpora*, 10(1):11–45.