Modelling grammatical alternation phenomena with aggregated and non-aggregated predictors

Felix Bildhauer (IDS Mannheim) and Roland Schäfer (Freie Universität Berlin)

Corpus studies on language varieties usually focus on differences in the distribution of linguistic features across different text categories. Among such categories, *genre*, *register* and *text type* play a very prominent role. However, more often than not, large corpora lack annotation for this kind of high-level categories, and the resources needed for manually generating such meta data are generally not available. Automatic annotation would seem to be the only viable alternative, and from the late 1980s on, there have been attempts at automatically classifying texts with respect to categories such as those mentioned above (see Karlgren and Cutting, 1994; Kessler et al., 1997; Lee and Myaeng, 2002; Freund et al., 2006; Kanaris and Stamatatos, 2009; Mehler et al., 2010, among others). These approaches usually (and plausibly) assume that relevant extra-linguistic characteristics of a text (e.g., its purpose) are reflected by certain linguistic features, and accordingly they exploit such features in the classification task. However, current experimental results from automatic genre detection in unrestricted domains are rather unsatisfactory (for example, Biber and Egbert, 2016, report 42.1% classification accuracy on 32 categories).

In our paper, we explore how useful automatically annotated text classes are when the focus is not on language varieties, but rather on syntactic and morphological alternation phenomena. More precisely, we examine the effects of aggregating linguistic features into a relatively small number of categories and using these categories to predict syntactic alternations, rather than predicting the alternation directly from the same linguistic features. To this end, we present two case studies of well-known alternation phenomena in German:

1. dative/genitive case alternations after prepositions, such as *wegen* 'because of', *gemäß* 'according to' and *einschließlich* 'including' (see, e. g., Di Meola, 2009)
2. inflection of adjectives after pronominal adjectives, such as *mit manch leckerem Kuchen* vs. *mit manchem leckeren Kuchen* vs. *mit manchem leckerem Kuchen* 'with some delicious cake' (see, e.g., Wiese, 2009)

Our corpus comprises 160,000 documents, about half of which sampled from the DECOW web corpus (Schäfer and Bildhauer, 2012; Schäfer, 2015) and the other half from the German Reference Corpus DeReKo (Kupietz et al., 2010). We automatically extract per-document counts for a number of linguistic categories (e. g., different parts-of-speech, morphological markers, syntactic constructions) as well as other features (such as text length, type/token ratio, emoticons). We then use the full set of these features as covariates in a generalized linear model (GLM), thus modelling directly the co-variance of these features with the alternation phenomenon. In the next step, we 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 genre/register/text type category. In a third GLM, only these document classes are used as predictors.

We compare the quality of the three 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 like genre, register, and text type in research on grammatical and morphological alternation phenomena.

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas and Egbert, Jesse. 2016. 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.

Di Meola, Claudio. 2009. Rektionsschwankungen bei Präpositionen - erlaubt, verboten, unbeachtet. In Marek Konopka and Bruno Strecker (eds.) (2009), 195–221.

Freund, Luanne, Clarke, Charles L. A. and Toms, Elaine G. 2006. Towards Genre Classification for IR in the Workplace. In *Proceedings of the 1st International Conference on Information Interaction in Context*, IIiX, 30-36, New York, NY, USA: ACM.

Kanaris, Ioannis and Stamatatos, Efstathios. 2009. Learning to Recognize Webpage Genres. *Information Processing and Management* 45(5), 499–512.

Karlgren, Jussi and Cutting, Douglass. 1994. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of COLING 94*, 1071–1075.

Kessler, Brett, Nunberg, Geoffrey and Schütze, Hinrich. 1997. Automatic Detection of Text Genre. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, ACL '98, 32–38, Stroudsburg, PA, USA: Association for Computational Linguistics.

Konopka, Marek and Strecker, Bruno (eds.). 2009. *Deutsche Grammatik - Regeln, Normen, Sprachgebrauch*. Berlin: de Gruyter.

Kupietz, Marc, Belica, Cyril, Keibel, Holger and Witt, Andreas. 2010. The German reference corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, 1848–1854, Valletta: ELRA.

Lee, Yong-Bae and Myaeng, Sung Hyon. 2002. Text Genre Classification with Genre-revealing and Subject-revealing Features. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, 145–150, NewYork, NY, USA: ACM.

Mehler, Alexander, Sharoff, Serge and Santini, Marina (eds.). 2010. *Genres on the web: computational models and empirical studies*, volume 42 of Text, speech and language technology. New York: Springer.

Roland Schäfer. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański et al. (eds.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, Lancaster: UCREL.

Schäfer, Roland and Bildhauer, Felix. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 486–493, Istanbul: ELRA.

Wiese, Bernd. 2009. Variation in der Flexionsmorphologie: Starke und schwache Adjektivflexion nach Pronominaladjektiven. In Marek Konopka and Bruno Strecker (eds.) (2009), pages 166–194.