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 leckerem 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
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.