The Register Casino — Should you risk your grammar in an outside bet?

Roland Schäfer (Freie Universität Berlin)

In this paper, I investigate the effects of using meta data generated from per-document distributions of text-internal features on the statistical modeling of grammatical alternation phenomena using data simulation. I argue that using lexical and grammatical features to classify texts according to *register* or *genre* and then including the output of this classification in the modeling of lexical or grammatical alternation phenomena is not only conceptually circular, but it is also detrimental to the quality of statistical models of so-called grammatical alternations. In the corpus-linguistic literature on grammatical alternation phenomena, Generalized Linear (Mixed) Models (GL[M]Ms; Gelman & Hill 2007) have become an accepted tool. Such models are used to quantify the influence of various features on a speaker's choice between two concurrent forms or two concurrent constructions (Gries 2014, 2015), such as the dative alternation in English (Bresnan et al. 2007) or the low-frequency alternation of the so-called weak nouns towards the strong inflectional pattern in German (Schäfer 2016, to appear). The underlying assumption is that sets of contextual features create a more or less prototypical environment for the variants, and the probabilities of either variant occurring in a specific context is modeled by the GLMM (also Divjak & Arppe 2013). The contextual features can be of virtually any kind, for example grammatical (such the case of an NP in a construction), semantic (such as humanness of a noun's denotation), pragmatic (givenness of a specific discourse referent), and even contextual (in a broader sense) or situational, such as the style, register, genre, or any similar category. It is often recommended that such information should be included as so-called *random effects* in GLMMs (Gries 2015).

Genre and register distinctions are usually not defined based on text-internal but text-external criteria (Lee 2001). However, if genre or register categorizations have to be made available for very large corpora (such as web corpora), they can only be achieved by automatic classification (see many contributions in Mehler et al. 2011). Since for practical reasons, automatic classification can only be based on features extractable from the documents themselves, only grammatical and lexical features can be used. This is very prominently true for Biber's (1995) register classification and Biber & Egbert's (2016) attempt at genre classification. Biber's register classification is a bottom-up procedure wherein per-text distributions of lexico-grammatical features are extracted and then reduced in dimensionality through factor analysis. The question is what happens when such register and genre classifications reconstructed from text-internal features are used as contextual features in GLMMs. Looking only at Biber's bottom-up register classification for now, I answer this question through data simulation. Data simulation and other *Monte Carlo* methods have gained prominence in many fields, including linguistics (Vasishth & Broe 2011; Carsey & Harden 2014). Data is artificially generated which has precisely specified properties assumed to be true for the relevant population. Properties of statistical procedures can then be quasi-experimentally examined through repeated sampling from the simulated population. The major advantages compared to real-life experiments are an infinite supply of data and the fact that we know the simulated population perfectly. This allows us to answer all kinds of "What if?" questions about the quality of statistical procedures under various circumstances – including negative effects of ill-formed data sets, suboptimal sampling procedures, faulty data aggregation, incorrect model specifications, etc. In my experiment, I simulated a population of texts and grammatical forms in these texts. In this population, I performed Biber-style multidimensional analysis using the text-level distribution of 40 grammatical features which were dimensionally reduced through factor analysis. Then, I simulated diverse grammatical alternations within the same population, estimating the corresponding GLMs (1) based on the raw features and (2) based on the register dimensions. Across the board, model quality drops significantly when the register dimensions were used:

prediction accuracy drops by 20% on average, R^2 by over 0.4. Furthermore, I can show that under many circumstances (e.g., multicollinearity), coefficient estimates become biased. I interpret the experiment as showing that we should either used truly externally defined (and necessarily *manually annotated*) document classifications or simply use raw measures of the distribution of grammatical features in our corpus studies. The reconstruction of register or genre from text-internal features is illusory, at least for the purpose of modeling grammatical alternation phenomena.

References

Biber, Douglas. 1995. *Dimensions of register variation*. Cambridge: Cambridge University Press.

Biber, Douglas & Jesse Egbert. 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.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Kraemer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Science.

Carsey, Thomas M. & Jeffrey J. Harden. 2014. *Monte Carlo Simulation and Resampling for Social Science.* Thousand Oaks: Sage Publications.

Divjak, Dagmar & Antti Arppe. 2013. Extracting prototypes from exemplars: What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). 221–274.

Gelman, Andrew & Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel / Hierarchical Models*. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2014. Corpus and quantitative methods. In Jeanette Littlemore & John R. Taylor (eds.), *The Bloomsbury companion to cognitive linguistics*, 279–300. London, New York: Bloomsbury.

Gries, Stefan Th. 2015. The most underused statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* 10(1). 95–126.

Lee, David. 2001. *Genres, registers, text types, domains, and styles: Claryfying the concepts and navigating a path through the BNC jungle*. Language Learning and Technology 5. 37–72.

Mehler, Alexander, Serge Sharoff & Marina Santini (eds). 2010. *Genres on the Web. Computational Models and Empirical Studies*. Dordrecht: Springer.

Schäfer, Roland. 2016, to appear. Prototype-driven alternations: the case of German weak nouns. To appear in *Corpus Linguistics and Linguistic Theory*.

Vasishth, Shravan & Michael Broe. 2011. *The Foundation of Statistics: A Simulation-based Approach*. Heidelberg etc.: Springer.