Automatic register annotation?

August 22, 2019

1 Introduction

Motivation: Variation, Biber, Bigbert & automatic register classification based on document-level counts of linguistic features; efforts to automatically generate document-level genre/register meta data based on such data. Aim: demonstrate more appropriate use of linguistic features in variational linguistic research; Outline: Present Biber-like feature extraction software for German; extract document-level features from corpora of newspaper and web texts; aggregate features via factor analysis; In 3 short case studies of (possibly register sensitive) morpho-syntactic variation phenomena, compare the "explanatory" power of the resulting factors to that of (selected) raw features, i.e., not using aggregation.

Clarifications:

1. The technical and statistical points made in this paper are not new; our aim is rather to demonstrate the impact of particular aggregation techniques in the context of variationist corpus linguist research, and discuss some of its consequences for corpus construction and corpus annotation.

2. We do *not* advocate "fishing", "snooping", "hunting": in practice, if the study is not explicitly declared as explorative, researchers should approach a given phenomenon with clear idea of which variables are relevant, based on "substantive theory".

3. The case studies presented here are merely intended to highlight the differences between models using aggregated data and models using non-aggregated data. So we focus on the automatically extracted linguistic features, deliberately neglecting other factors that can be expected to play a role in modeling a given phenomenon.

Figure 1: The COReX feature extractor

2 The COReX feature extractor

COReX is a piece of software designed to extract a large number of normalised feature counts[1] at the document level from German text. Most features correspond to linguistic concepts (such as occurrences of a particular parts-of-speech, periphrastic passive and perfect constructions, non-standard morphological forms), alongside a few features which are not strictly linguistic (such as type/token ratio and word/sentence lengths). COReX does not perform linguistic annotation by itself, but rather relies on linguistic pre-processing, typically performed automatically by dedicated annotation tools. COReX is implemented in Python and is easily extendable to include additional features (which, in most cases, will require additional pre-processing of the input corpus).[2]

In order to extract the full range of features, the data must include part-of-speech tags (STTS; Schiller et al. 1999), morphological features (such as those produced by MarMoT; Müller, Schmid & Schütze 2013), named entity annotations (such as those produced by the Stanford NER tagger; Finkel, Grenager & Manning 2005) as well as topological field annotations (as produced by the Berkeley parser; Petrov & Klein 2007, Cheung & Penn 2009, Telljohann et al. 2012).

[Maybe one or two plots illustrating the distribution of some selected features (e.g., crx_gen and crx_prep; or some plots from k-medoid clustering, but this is maybe too much]

[1] Currently, COReX extracts over 60 features. The complete list is provided in the appendix.
[2] Another note.

3 Feature aggregation

Information derived from counts of linguistic features in texts can be used in various ways for characterizing the the texts they were extracted from. In pioneering work by Biber 1988, Biber 1995, factor analysis was used to identify a number of "dimensions of variation" that could be interpreted in terms of communicative function (such as "narrative vs. non-narrative", "involved vs. informational production"), and individual texts can then be located at different points on each one of these scales. Contrasting with this multidimensional approach, there have also been attempts to automatically assign texts to single a category ("genre", "text type") based on the occurrence of linguistic features on those texts. An early example of such an approach is Karlgren & Cutting 1994. A recent state-of-the-art example is reported in Egbert, Biber & Davies 2015, which illustrates the enormous challenge of automatic register classification on the basis of linguistic feature counts. Using 44 linguistic features in a discriminant analysis for predicting the register category of documents from an "unrestricted corpus of web documents", they report precision = 0.342 and recall = 0.396 for their 20 specific sub-registers used in the task (results are generally lower when a smaller number of broader, less specific register categories is used).

In order to explore the effects of feature aggregation in predicting potentially register-sensitive morpho-syntactic alternation phenomena, we use a multidimensional approach (based on Factor Analysis) as described in Biber 1988, 1995. In Section 4, we will treat the documents' factor scores on each one of the resulting factors as document meta data, and use that meta data in modeling the outcome in specific instances of the alternation phenomenon. Note that in using a multidimensional approach, we still produce several document-level predictors. In contrast, assigning each document to a single register category would produce only a single document-level predictor, leading to a greater loss of information.

3.1 Factor analysis

In general, we sought to reproduce the technical aspects of Biber's (1988) study as closely as possible. There is a major difference, however, in corpus size. While Biber used material from a total of 481 documents, we sampled 70,000 documents from DeReKo and DECOW16B each, totaling 140,000 documents. Another difference concerns document length. Biber 1988 imposes an upper limit of approximately 2,500 words per document, whereas we base our counts on whole documents, without an upper limit to the number of words. The minimum text length is 400 words in Biber's study. In contrast, we use a lower threshold of only 100 tokens (many documents DeReKo are actually shorter than that).

For the factor analysis, we used a total of 60 feature counts extracted with COReX.[3] The *genitive* count, originally extracted by COReX, was discarded in view of the case study reported in Section 4, where the occurrence of genitive is predicted. All normalized counts were scaled to z-scores. Visual inspection of a parallel analysis scree plot suggested an optimal number of 7 to 8 factors for our data set. The plot in Figure 3.1 shows the factor loadings from a factor analysis using 7 factors. The factoring and rotation methods (*principal factor* and *promax*, respectively) were chosen to match those used and recommended in Biber 1988.

Some, but not all of the factors lend themselves to an interpretation in terms of meaningful dimensions of variation. Most notable among these is Factor 1. Features with high factor loadings on Factor 1 include short/ contracted forms, interjections, emoticons, imperatives, vocabulary typical of informal written language, as well as first and second person pronouns. Factor 1 thus most probably captures variability along the lines of formal vs. informal, standard vs. non-standard, high vs. low degree of interaction. On closer inspection of a number of documents with high scores on Factor 1, it turned out that "non-standard" also extends to dialectal texts. In contrast, finding a plausible interpretation for factors such as Factor 2 is not as straightforward. Features positively associated with Factor 2 include finite verbs, lexical verbs, auxiliary verbs and third person pronouns, while a significant number of features are negatively associated with it (including cardinal numbers and various types of named entities).

Factor scores were computed for every factor for each one of the 140,000 documents in the sample, along the lines of Biber 1988, p. 93–97, as follows:

- Factor loadings whose absolute value is less than 0.35 were ignored.
- Any given feature is part of only one factor (the one with the greatest loading for that feature).
- For each document, for each factor, standardized counts (i. e., z-scores) were added up for those features that are salient on that factor (salient meaning the feature's absolute loading is at least .35 and there is no other factor where this feature has a greater absolute loading).

This approach implies that the exact magnitude of a feature's loading on a factor is irrelevant for the calculation of a document's factor scores; for instance, it does not make a difference whether a loading is .35 or .99.

Biber 1988 proceeds by comparing the mean factor scores of documents belonging to different (externally defined) registers (such as *romantic fiction*, *biographies*, *press reviews*) on various dimensions of variation (= interpreted factors).

[3] Absolute counts, such as number of tokens and number of sentences, were discarded.

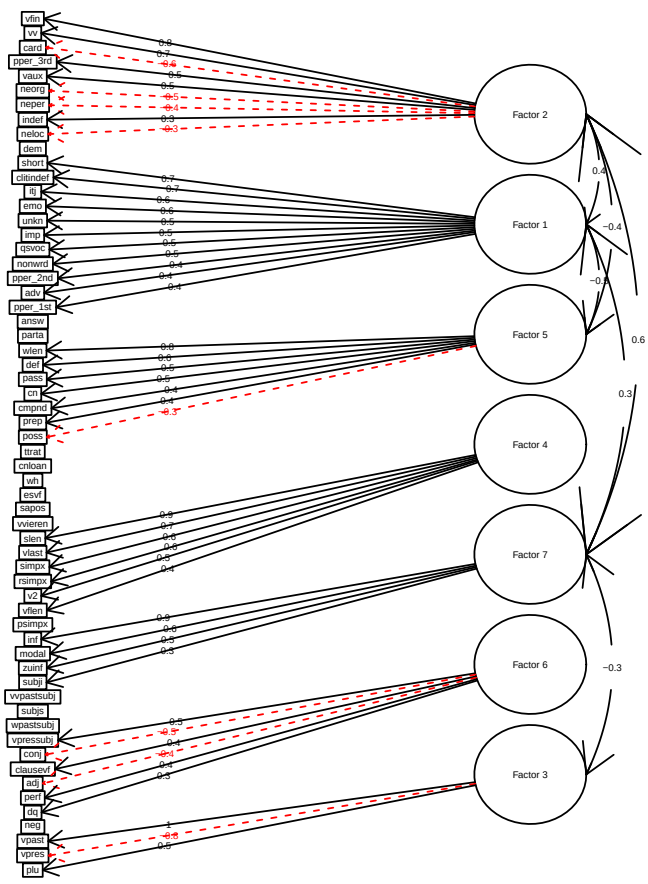

Figure 2: Factor loadings of 60 COReX features, from a factor analysis of 140.000 randomly selected documents (DeReKo and DECOW16 combined, minimal document length = 100 tokens), 7 factors, principal factor method, promax rotation. Only the highest factor loading is shown for each feature (loadings between −0.35 and 0.35 are omitted).

Unfortunately, as was discussed above, the corpora used in our study hardly contain any reliable register information at all. One distinction that can be made, though, is between texts from internet discussion forums (recognizable by their URL patterns) and any texts (web or other) that are not forum discussions. We would expect forum discussions to exhibit a high degree of interaction (question/comment - reaction/answer), as well as a fair amount of non-redacted, non-standard language, and this is indeed what we find when we compare the distribution of document scores on Factor 1 for forum (mean 14.6) and non-forum (mean -1.3) documents, as shown in Figure 3.1.

By generating factor scores for each of the 7 factors for each document, we have enriched our corpus with 7 additional meta data variables at the document level. In the next section, we will use these variables as document-level predictors in a study on a specific syntactic alternation phenomenon, and compare the results to alternative models that use non-aggregated COReX counts as predictors.

4 Case studies

In this section, we explore the consequences of aggregating individual linguistic predictors into more abstract factors, for the purpose of modeling linguistic alternation phenomena. In a series of case studies, we model particular morpho- series?
syntactic alternation phenomena with generalized linear models (GLMs), using only document-level information. Each time, we use as predictors the 7 factor scores from the factor analysis described in the previous section. In addition, we specify an alternative model, directly using the set of COReX features as predictors, and we compare these models wrt. to model fit and prediction accuracy. We are aware that these linguistic phenomena could probably be better explained / modeled if other kinds of predictors were taken into account as well (e. g., lexical information, syntactic properties at various levels). However, since we are interested in what different sorts of document-level information can contribute to modeling the alternation phenomena, we deliberately ignore other predictors as part of our study design.

4.1 Dative vs genitive case governed by prepositions

A number of prepositions in German show variation in assigning case to their NP complement. A well-known example is the accusative/dative alternation after certain prepositions, which systematically encodes a semantic distinction (directional/non-directional movement; REF). Contrasting with this kind of semantically relevant alternation, some prepositions exhibit variation in case assignment which is not semantically motivated, but rather considered stylistic. The present case study

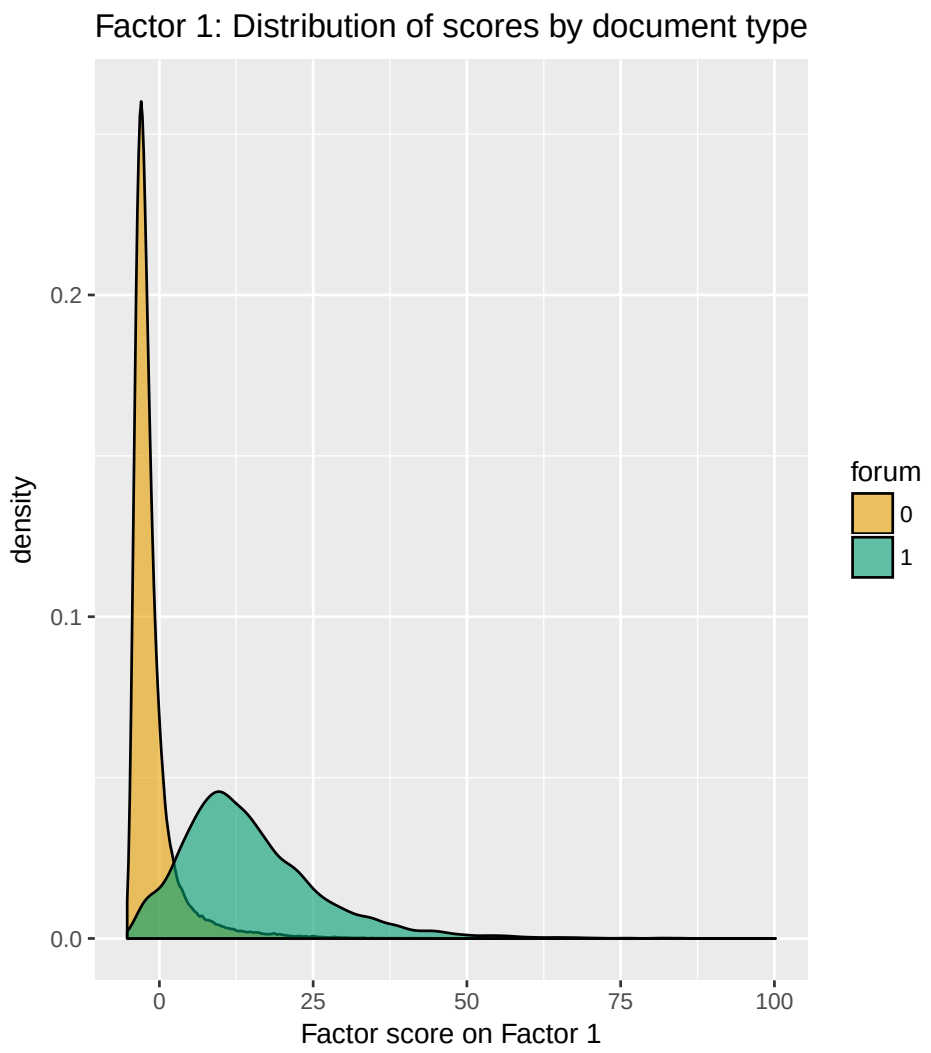

Figure 3: Distribution of factor scores (Factor 1). On average, forum documents show higher scores (14.6) than non-forum documents (−1.3). Prominent features on Factor 1 include short/contracted word forms, cliticised variants of the indefinite article, interjections, imperatives, 1st and 2nd person pronouns, emoticons, and words typical of informal written language.

focuses on alternations of the second type, involving genitive and dative case, as illustrated in (1)–(4).

(1) a. trotz [starkem Verkehr]$_{dat}$
'despite heavy traffic'
b. trotz [ihres Namens]$_{gen}$
'despite her name'

(2) a. wegen [dem Geschmack]$_{dat}$
'because of the taste'
b. wegen [des besseren Aussehens]$_{gen}$
'because of the better appearance'

(3) a. entgegen [dem ursprünglichen Gesetzentwurf]$_{dat}$
'contrary to the original bill'
b. entgegen [des Gesamttrends]$_{gen}$
'contrary to the overall trend'

(4) a. gegenüber [einem Dritten]$_{dat}$
'vis-à-vis a third party'
b. gegenüber [des Hotels]$_{gen}$
'opposite the hotel'

Typically, one of the variants is considered as normative / canonical in Standard German, while the competing form has in many cases a non-standard flavour (see e. g., Di Meola 2009 and references therein).[4] We therefore expect that the choice of case after these prepositions depends partially on register and, more specifically, on the dimension of variation captured by Factor 1. Such case alternations thus provide a promising area for exploring the effects of feature aggregation in modeling register-sensitive linguistic alternation phenomena.

For the present case study, we selected a number of prepositions likely to exhibit some variation in dative/genitive case:

> abzüglich, angesichts, anlässlich, außer, betreffs, bezüglich, dank, einschließlich, entgegen, gegenüber, gemäß, hinsichtlich, mangels, mitsamt, mittels, nebst, samt, seitens, trotz, vorbehaltlich, während, wegen, zuzüglich

[4]For some prepositions, the normative case depends on morpho-syntactic properties of the complement NP. For instance, *trotz* canonically assigns genitive, but dative is the only acceptable option with bare plural nouns which would otherwise lack a genitive inflectional ending. In what follows, we will only consider syntactic contexts where speakers/writers actually have a choice between genitive and dative.

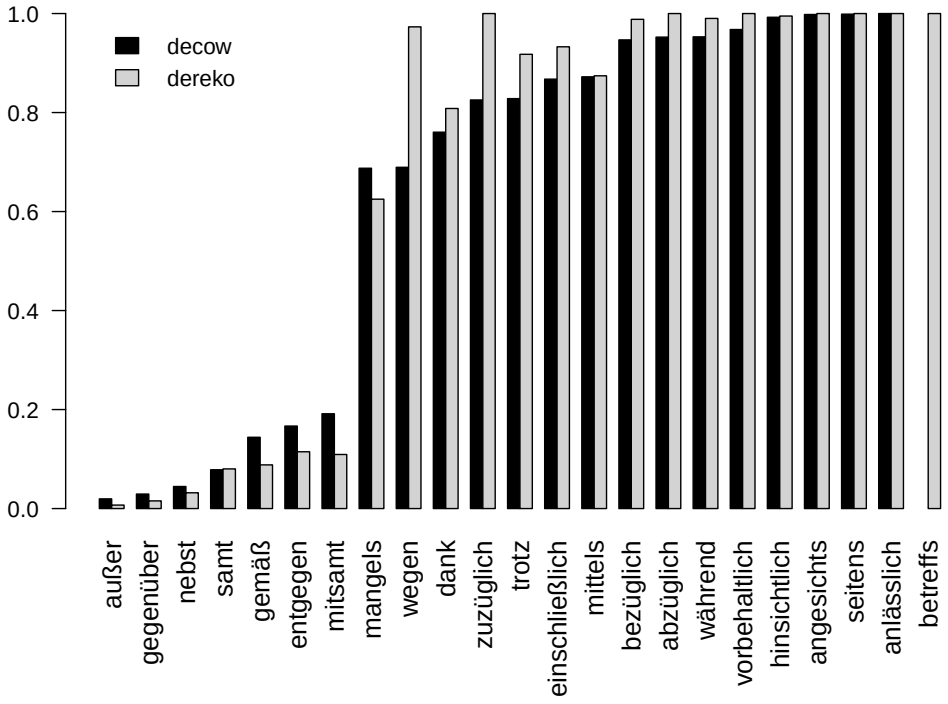

Figure 4: Proportion of genitive complements (as opposed to dative) by preposition and corpus, in a sample of 80,000 prepositional phrases from DeReKo and DECOW16B

A fair number (but by no means all) of the selected prepositions show some degree of variation in case assignment. For our final dataset, we selected only those prepositions where the proportion of either genitive or dative occurrences was between 0.1 and 0.9 in at least one of the corpora. In other words, the amount of variation must be such that at least 10% of all occurrences of that preposition show the minority, non-modal (and arguably, non-standard) case category in at least one of the corpora. By this criterion, only 10 out of the original 23 prepositions were included in the final dataset (*entgegen*, *gemäß*, *mitsamt*, which typically select a dative complement, as well as *dank*, *einschließlich*, *mangels*, *mittels*, *trotz*, *wegen* and *zuzüglich*, which typically select genitive). For this dataset, the overall occurence rate of non-standard case assignment is 15.4 %.

We first specify two logistic regression models that predict the probability of observing the non-modal/non-standard case category, and which do not distinguish between individual prepositions. First, we use the set of COReX variables, represented as $c_1 \ldots c_{60}$ in equation 5.

Variable name

$$P(nonstandard.case = 1) = logit^{-1}(\alpha + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_{60} c_{60}) \quad (5)$$

Of the resulting coefficient estimates, 32 are different from 0 at $p < 0.05$. The Nagelkerke Pseudo-R^2 score for this model is 0.28.

For comparison, we use the document factor scores from the factor analysis as shown in equation 6 (where the terms $f_1 \ldots f_7$ represent the factor scores of factors 1 through 7). In this model, all coefficient estimates are significant at the 0.05 level, however, the Nagelkerke R^2 score for this model drops to 0.24.

Matrizenoder Laufindex oder Fußnote und erklären

$$P(nonstandard.case = 1) = logit^{-1}(\alpha + \beta_1 f_1 + \beta_2 f_2 + \ldots + \beta_7 f_7) \quad (6)$$

Next, we consider separate models for each preposition, using all 60 COReX features as predictors. Figure 4.1 illustrates the distribution of estimates for each preposition, for each coefficient with an associated p-value < .05 and absolute value < 5. As is obvious from the plot, coefficient estimates vary greatly, depending on the preposition.

References

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press. http://www.loc.gov/catdir/toc/cam023/87038213.html.

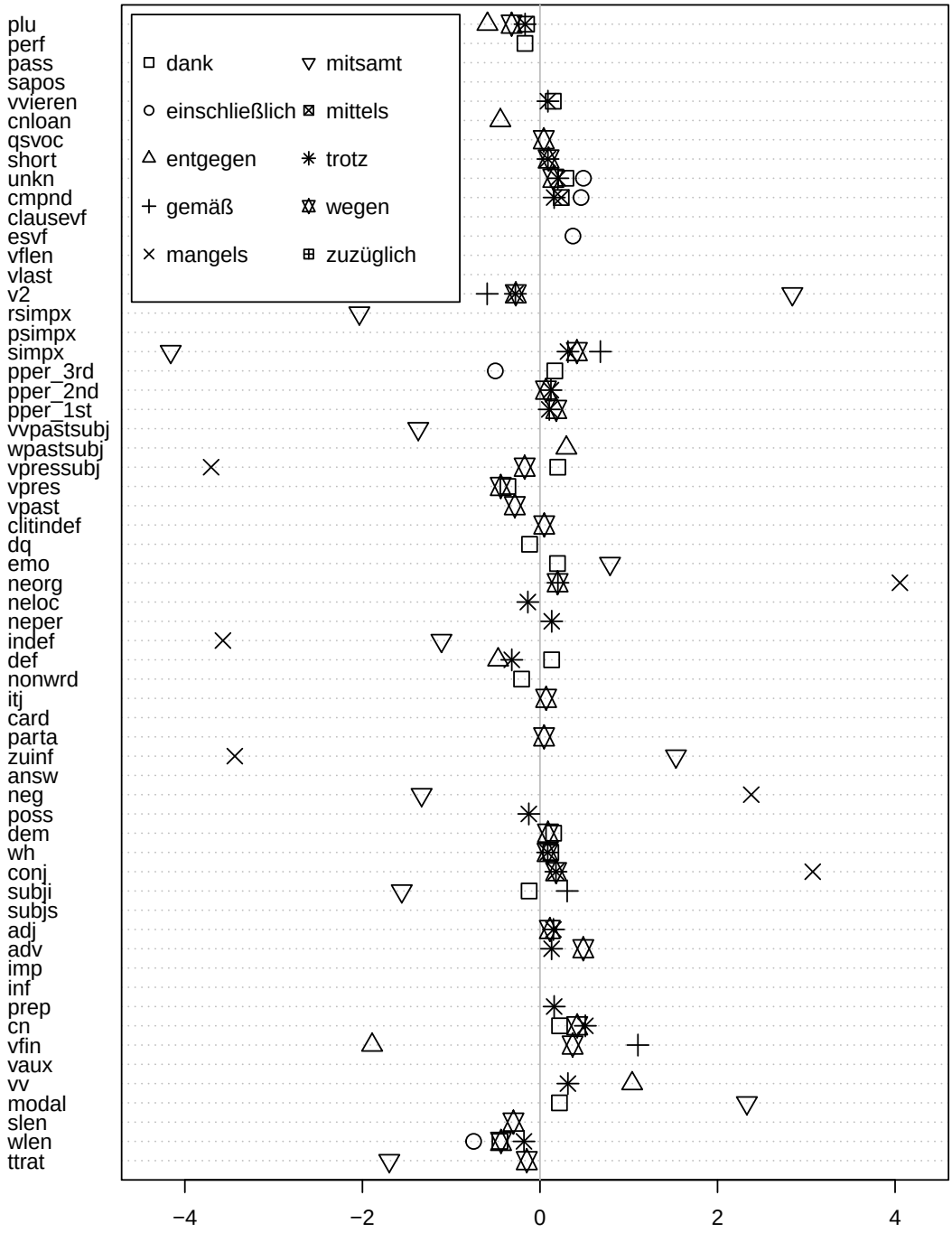

Figure 5: COReX features: coefficient estimates with associated p-value < 0.05; a separate model was specified for each preposition.

Biber, Douglas. 1995. *Dimensions of register variation: a cross-linguistic comparison.* Cambridge: Cambridge University Press. http://www.loc.gov/catdir/toc/cam023/94017041.html.

Cheung, Jackie Chi Kit & Gerald Penn. 2009. Topological field parsing of german. In *Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the afnlp*, 64–72. Suntec, Singapore: Association for Computational Linguistics.

Di Meola, Claudio. 2009. Rektionsschwankungen bei präpositionen - erlaubt, verboten, unbeachtet. In Marek Konopka & Bruno Strecker (eds.), *Deutsche grammatik - regeln, normen, sprachgebrauch*, 195–221. Berlin etc.: de Gruyter.

Egbert, Jesse, Douglas Biber & Mark Davies. 2015. Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology* 66(9). 1817–1831. http://dx.doi.org/10.1002/asi.23308.

Finkel, Jenny Rose, Trond Grenager & Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43nd annual meeting of the association for computational linguistics (acl 2005)*, 363–370. Association for Computational Linguistics.

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

Müller, Thomas, Helmut Schmid & Hinrich Schütze. 2013. Efficient higher-order CRFs for morphological tagging. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 322–332. Seattle, Washington, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/D13-1032.

Petrov, Slav & Dan Klein. 2007. Improved inference for unlexicalized parsing. In *Human language technologies 2007: the conference of the north american chapter of the association for computational linguistics; proceedings of the main conference*, 404–411. Rochester, New York: Association for Computational Linguistics.

Schiller, Anne, Simone Teufel, Christine Stöckert & Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textkorpora mit STTS (kleines und großes Tagset).* Institut für maschinelle Sprachverarbeitung, Universität Stuttgart & Institut für Sprachwissenschaft, Universität Tübingen. Stuttgart & Tübingen.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister & Kathrin Beck. 2012. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).* Tech. rep. Universität Tübingen Seminar für Sprachwissenschaft.