Automatic register annotation?

December 7, 2018

1 Introduction

Motivation: Variation, Biber, Bigbert & automatic register classification based on docuement-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.

Clarification #1: 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”.

Clarification #2: 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.

2 The COReX feature extractor

COReX is a piece of software that extracts a large number of normalised linguistic feature counts at the document level from German text.1 It does not perform any linguistic annotation by itself, but instead requires linguistic pre-processing, typically performed automatically by dedicated annotation tools. The input format is vertical text (one token per line) with token level annotations in columns 2 through n, and token spans represented as XML-elements (this corresponds to the input format required by tools such as OpenCWB and NoSketchEngine). Required XML-elements are doc (document) and s (sentence). In order to extract the full range of features, the data must include part-of-speech tags (STTS,), morphological features (such as those produced by MarMoT,Müller et al., 2013), named entity annotations (such as those produced by the Stanford NER tagger, Finkel et al., 2005) as well as topological field annotations (Berkeley, REF). For some annotation layers, such as morphological features and topological fields, a particular formatting is required. COReX outputs a modified version of the input data where normalised document-level feature counts are added as attribute-value pairs to the opening <doc > tag for each document. Moreover, a number of non-normalised counts (such as perfect and passive) are added as attribute-value pairs to the <s > tag for each sentence. Features include non-linguistic categories such as mean word length and mean sentence length; counts for individual parts-of-speech; ...

[Distribution of features in COW data; clustering]

3 Aggregation: factor analysis

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, 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 a category ("genre", "text type") based on the occurrence of linguistic features on those texts. A recent state-of-the-art example of such an approach is reported in Egbert et al. (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

Factor Analysis as described in Biber (1988, 1995). In Section , we will treat the documents' factor scores on each one of the resulting factors as document meta, and use that meta data for modelling the outcome in specific instances of the alternation phenomenon.

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-syntactic alternation phenomena with generalized linear models (GLMs), using only document-level information. Each time, we use as predictors the full set of linguistic feature counts as extracted by COReX. In addition, we specify an alternative model using as predictors the per-document factor loadings from a factor analysis. Finally, we compare these models wrt. to model fit and prediction accuracy. We are aware that some or all of the phenomena could probably be better explained / modelled 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 modelling 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 wrt. case marking of their NP complement. A well-known example is the accusative/dative alternation after certain prepositions, which systematically encodes a semantic distinction (directional/non-directional; 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 focuses on the alternation between 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'

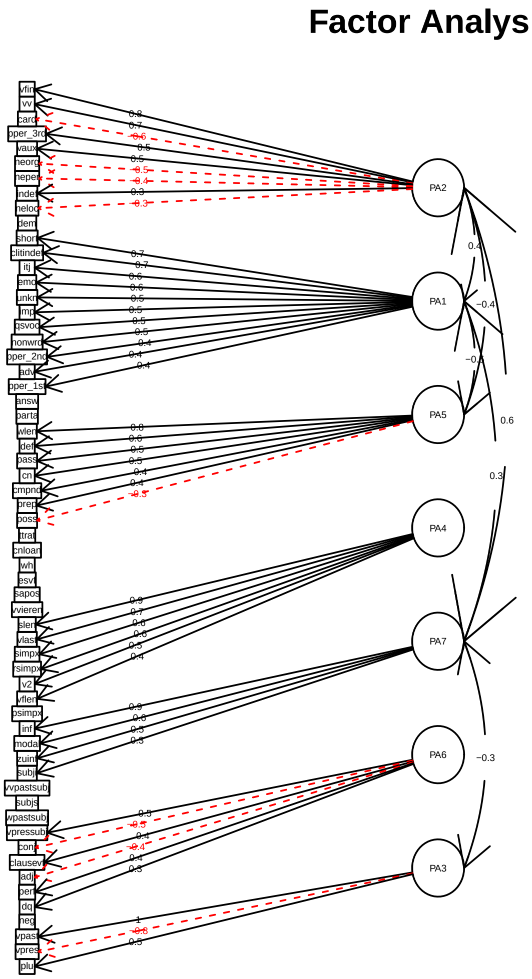

Figure 1: Factor loadings for COReX features; Factor analysis of 140.000 random documents (DeReKo and DECOW16 combined, minimal document length = 100 tokens), 7 factors, principal factor method, promax rotation

b. wegen [des besseren Aussehens]$_{gen}$
'because of the better appearance'

(3) a. entgegen [dem ursprünglichen Gesetzentwurf]$_{dat}$
'contrary to the original draft 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 regarded 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). We therefore expect that the choice of case after these prepositions will depend, at least in part, on text type / genre, thus providing an ideal area for exploring the effects of feature aggregation. 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

We used the German web corpus DECOW16B (Schäfer and Bildhauer, 2012) as well as the subset of the German reference corpus DeReKo (Kupietz et al., 2010) documented inBubenhofer et al. (2014). For each of these prepositions, all occurrences were extracted where the preposition is followed by either a determiner or an adjective that unambiguously mark dative or genitive case (cf. examples 1 and 8 above). Concordance lines from documents containing less than 100 tokens were discarded, so as to ensure that document-level feature counts are reasonably reliable. Moreover, we only kept a single instance per document, discarding all remaining instances. From this preliminary sample, we randomly selected 40,000 instances per

[0] 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 genitive and dative.

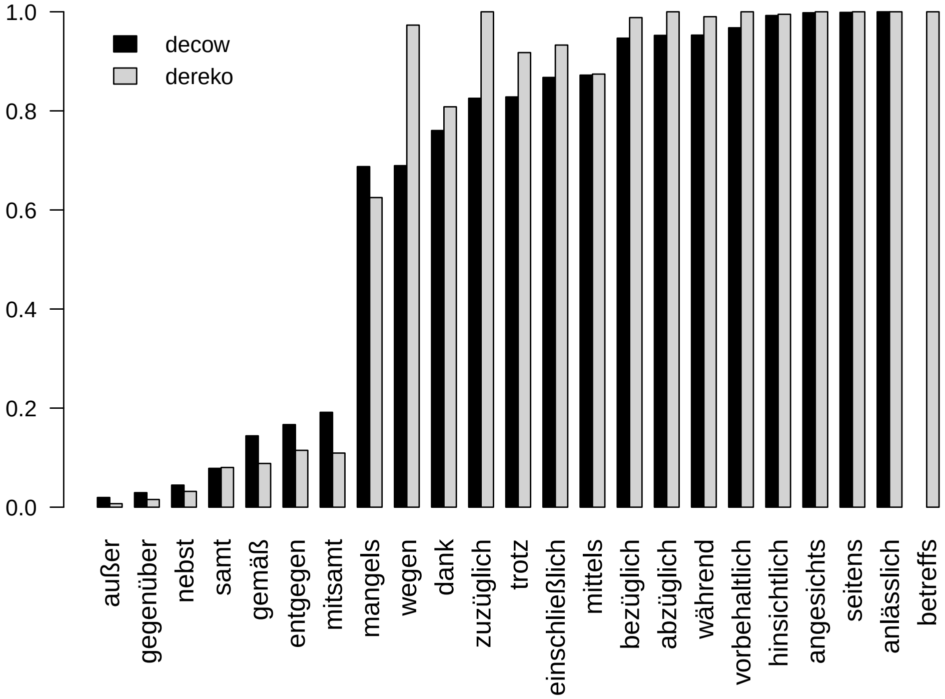

Figure 2: Proportion of genitive complements (as opposed to dative) by preposition and corpus, in a sample of 80,000 prepositional phrases from DeReKo and DECOW16B

corpus. Figure 4.1 shows the proportion of genitive complements by preposition and corpus.

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 ocurrences 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. Thus, from among the original 23 prepositions, only 10 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 1.

Variable name

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

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 2 (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

$$P(\mathit{nonstandard.case} = 1) = \mathit{logit}^{-1}(\alpha + \beta_1 f_1 + \beta_2 f_2 + \ldots + \beta_7 f_7) \quad (2)$$

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.

Warum geht das nicht: Biber (1988)

References

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

Biber, Douglas. 1995. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge: Cambridge University Press.

Bubenhofer, Noah, Konopka, Marek and Schneider, Roman. 2014. *Präliminarien einer Korpusgrammatik*, volume 4 of *Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache*. Tübingen: Narr, unter Mitwirkung von Caren Brinckmann.

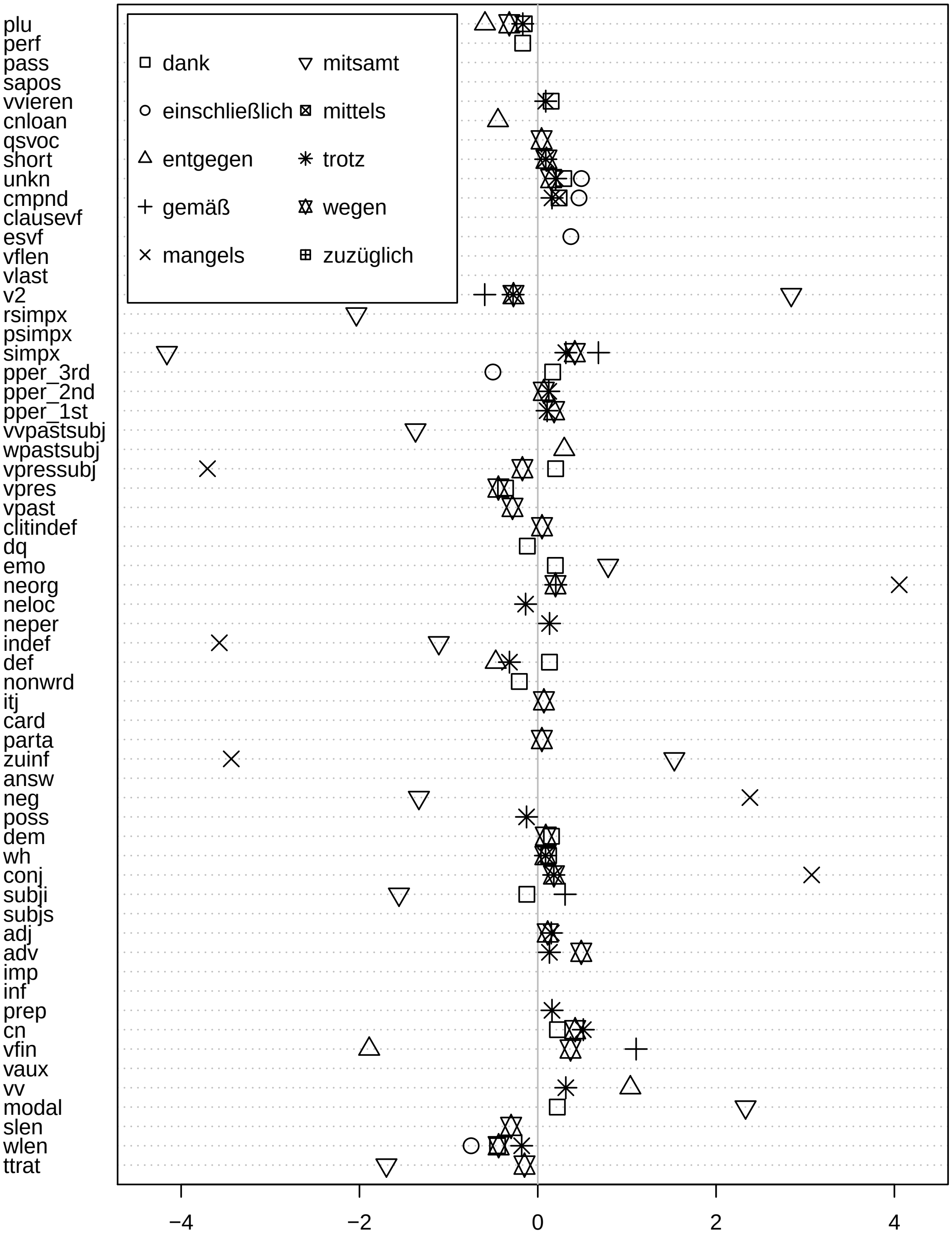

Figure 3: COReX features: coefficient estimates with associated p-value < 0.05; a separate model was specified for each preposition.

Di Meola, Claudio. 2009. Rektionsschwankungen bei Präpositionen - erlaubt, verboten, unbeachtet. In Marek Konopka and Bruno Strecker (eds.), *Deutsche Grammatik - Regeln, Normen, Sprachgebrauch*, pages 195–221, Berlin etc.: de Gruyter.

Egbert, Jesse, Biber, Douglas and Davies, Mark. 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.

Finkel, Jenny Rose, Grenager, Trond and Manning, Christopher. 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)*, pages 363–370, Association for Computational Linguistics.

Kupietz, Marc, Belica, Cyril, Keibel, Holger and Witt, Andreas. 2010. The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC '10)*, pages 1848–1854, Valletta, Malta: European Language Resources Association (ELRA).

Müller, Thomas, Schmid, Helmut and Schütze, Hinrich. 2013. Efficient Higher-Order CRFs for Morphological Tagging. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 322–332, Seattle, Washington, USA: Association for Computational Linguistics.

Schäfer, Roland and Bildhauer, Felix. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk and Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul: ELRA.