# Automatic register annotation?

June 9, 2020

## 1 Alternation modelling and register classification

### 1.1 Corpus-based modelling of linguistic variation

Modelling linguistic variation in the broadest sense is clearly one of the major goals – if not *the* major goal – of linguistics. Speakers trivially do not use the same expressions for non-identical content (compositional semantics) or communicative intent (pragmatics). They do not use the same expressions even for the same content or communicative intent in different regions (regio- or dialect), different social groups (sociolect), or based on individual preference (style). Furthermore, speakers use different forms in different communicative settings (such as talking to close personal friends vs. giving a talk in the process of applying for a tenured academic position), which is what we understand as *register*. Linguistic variation – even under a fairly traditional view – thus occurs at the level of individual utterances (driven by semantics and pragmatics), the level of individual speakers (style), the level of groups of speakers (dialect and sociolect), and the within-speaker level (style). Furthermore, usage-based approaches, which have been on a powerful rise for fifteen to twenty years now, stress that the distribution of concrete forms in utterances is also driven partially by input co-occurrence frequencies. Under this view, variation is also biased by the fact that some forms happen to be more frequently used together than others, be it for obvious functional or less transparent or obscured reasons. Speakers most likely also chose their forms partially based on such co-occurrence preferences.[1] Finally, random variation should also be expected, either as an effect of so-called *performance* (i. e., production errors based on processing limitations) or because the linguistic system itself (at least as represented in the brains of speakers) does not represent a discrete (traditional linguistic competence) but rather a stochastic system (probabilistic grammar).[2] Modelling linguistic variation thus subsumes the task of specifying why speaker uses the specific expressions that they use under any given circumstances. Thus, if all sources of variation had been pinpointed and modelled, a full model of human language would emerge under this broad view.

[1]It is clear at this point that all dimensions of variation might overlap and be difficult to disentangle. Also, variation based on co-occurrence frequencies does not represent an independent dimension of variation as long as there is a transparent functional motivation for a co-occurrence preference. However, the usage-based perspective adds a stochastic and input-based tone to the picture.

[2]Luckily, the heated debate between those two views of random variation in linguistic output is not relevant to this study.

## 1.2 Potentials of incorporating register

## 1.3 The problems with registers in corpora

## 1.4 Statistical issues

## 1.5 Goals and overview

vgl. S. Müller 2010, pp. 33–36


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[3] 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).[4]

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; T. 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]

## 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

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

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[5] 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.[6] 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).

[5]More precisely, we used a stratified DeReKo subset of approx. 7 bn tokens, as described in **Bubenhofer-ea2014**, and which was annotated much in the same way as DECOW16B.

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

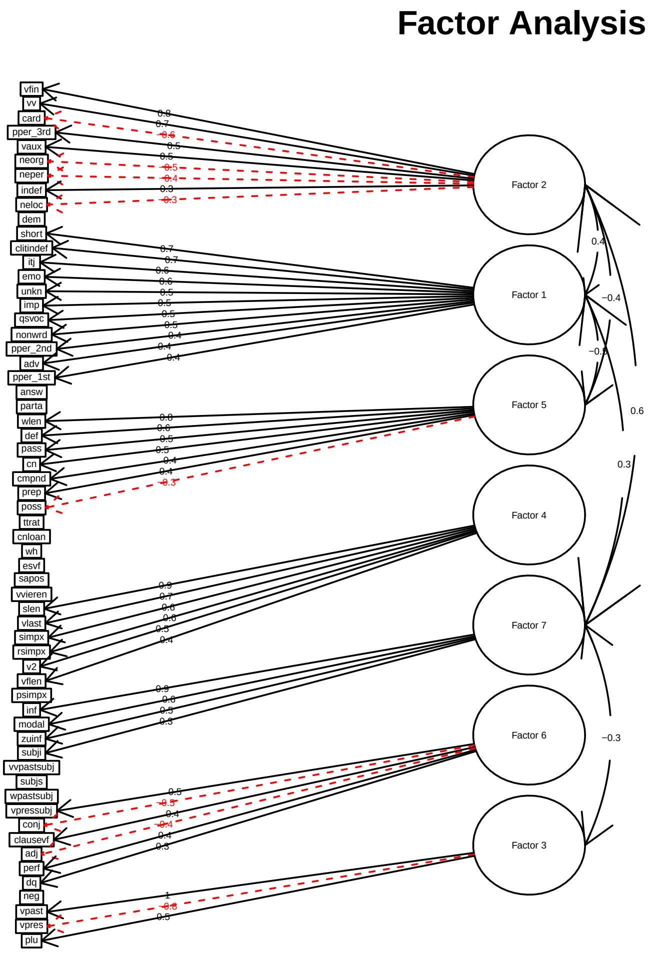


Figure 2: Factor loadings of 60 COReX features, from a factor analysis of 140.000 randomly selected documents (DeReKo and DECOW16 combined, minimal document length = 100 tokens), 7 factors, principal factor method, promax rotation. Only the highest factor loading is shown for each feature (loadings between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

- 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). 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-syntactic alternation phenomena with generalized linear models (GLMs), using different sorts of document-level information as predictors. For each case, we specify three alternative models, estimate the model coefficients, and compare these models wrt. to model fit and prediction accuracy: series?

1. A model which uses as predictors the set of 60 individual COReX features.
2. A model which uses as predictors the seven factor scores from the factor analysis described in the previous section.
3. A model which uses as predictors the same seven factor scores as model 2, plus another 53 predictor variables containing random data from a normal distribution (mean = 0, sd=1). This is to allow for a fair comparison with the COReX-based models (1), which otherwise could show better model fit merely because of accidental correlations between some of the many COReX features and the outcome.

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.

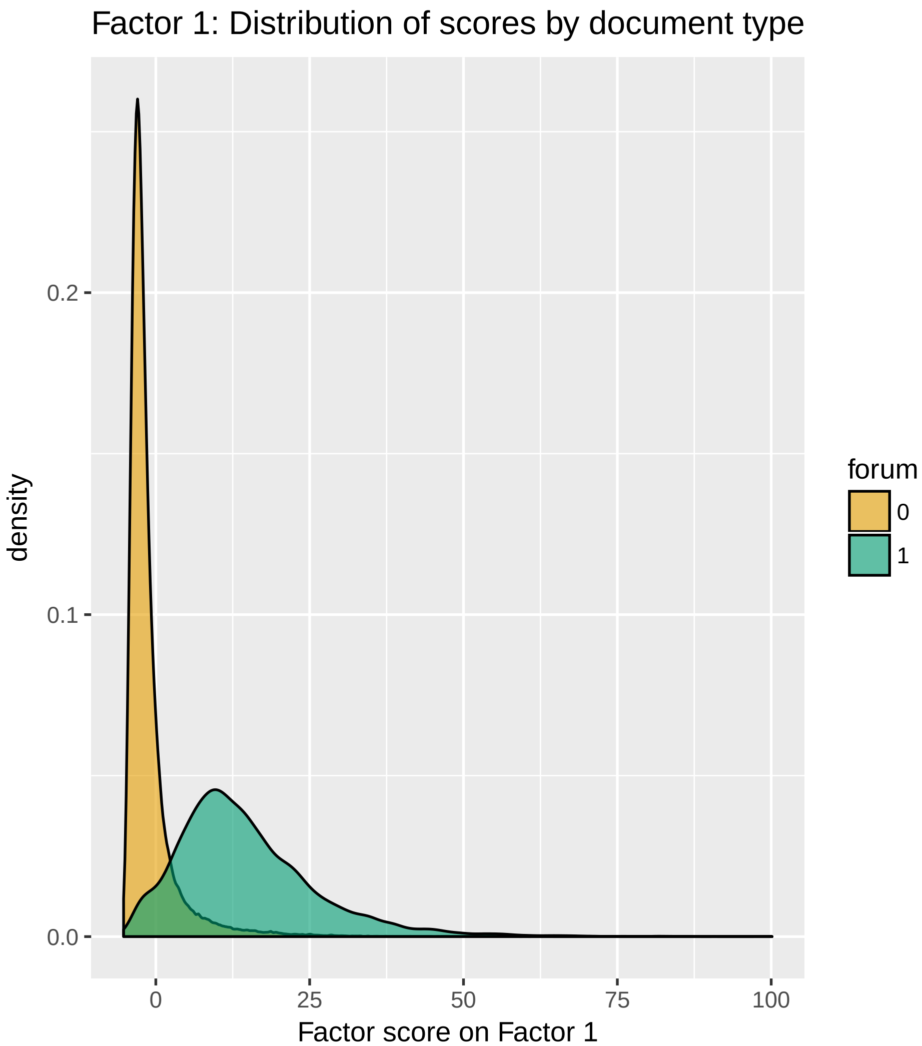


Figure 3: Distribution of factor scores (Factor 1). On average, forum documents show higher scores (14.6) than non-forum documents (−1.3). Prominent features on Factor 1 include short/contracted word forms, cliticised variants of the indefinite article, interjections, imperatives, 1st and 2nd person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

## 4.1 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 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).[7] 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

Occurrences of these prepositions were extracted from the DeReKo and DECOW16B corpora, restricted to contexts where case (either dative or genitive) is unambiguously marked on the prepositions's complement. More precisely, we considered sequences of the form

1. PREPOSITION – ADJECTIVE – NOUN (singular, masc./neut.)
   *dank soliden Lebens*
   *wegen dringendem Tatverdacht*

[7]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.

| Preposition | nscase.decow | nscase.dereko | n.decow | n.dereko | n.total |
|---|---|---|---|---|---|
| dank | 0.26 | 0.20 | 3786 | 3776 | 7562 |
| einschließlich | 0.13 | 0.08 | 3767 | 3773 | 7540 |
| entgegen | 0.16 | 0.10 | 3747 | 3726 | 7473 |
| gemäß | 0.15 | 0.10 | 3704 | 3790 | 7494 |
| mangels | 0.34 | 0.24 | 3707 | 2311 | 6018 |
| mitsamt | 0.13 | 0.09 | 3728 | 3687 | 7415 |
| mittels | 0.13 | 0.13 | 3742 | 3716 | 7458 |
| trotz | 0.17 | 0.08 | 3781 | 3732 | 7513 |
| wegen | 0.32 | 0.03 | 3713 | 3612 | 7325 |
| zuzüglich | 0.23 | 0.14 | 3351 | 859 | 4210 |

2. PREPOSITION – ARTICLE WORD[8]
   *wegen des*
   *gegenüber dem*

Many, though not 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 each of these, we randomly selected a maximum of 4,000 occurrences per corpus, and kept only a single instance per document in case there were more than one. The final data set is summarized in Table 4.1. Figure 4.1 shows the proportion of non-standard case in this dataset, by preposition and corpus.

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.

Matrizen oder 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 first 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.

[8]For the purpose of this study, article words include the definite and indefinite article ((*der*, *ein*), demonstratives (*dieser*, *jener*), possessives (e. g., *mein*, *unser*) as well as *irgendein*.

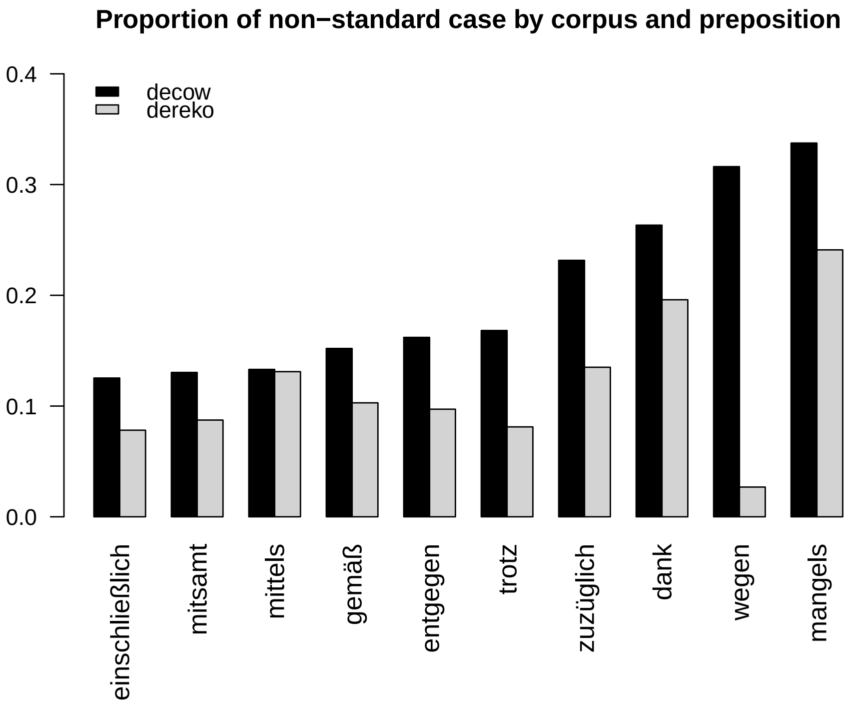


Figure 4: Proportion of non-standard case by preposition and corpus, in a sample of 70,008 prepositional phrases from DeReKo and DECOW16B

| Preposition | Total | nscase | p(nscase) | $R^2_{COReX}$ | $R^2_{FA}$ | $R^2_{FA+Noise}$ |
|---|---|---|---|---|---|---|
| dank | 7562 | 1737 | 0.230 | 0.085 | 0.046 | 0.055 |
| einschließlich | 7540 | 767 | 0.102 | 0.112 | 0.061 | 0.078 |
| entgegen | 7473 | 969 | 0.130 | 0.067 | 0.028 | 0.041 |
| gemäß | 7494 | 953 | 0.127 | 0.028 | 0.003 | 0.020 |
| mangels | 6018 | 1808 | 0.300 | 0.214 | 0.135 | 0.147 |
| mitsamt | 7415 | 808 | 0.109 | 0.043 | 0.007 | 0.017 |
| mittels | 7458 | 985 | 0.132 | 0.088 | 0.037 | 0.056 |
| trotz | 7513 | 939 | 0.125 | 0.128 | 0.087 | 0.101 |
| wegen | 7325 | 1271 | 0.174 | 0.503 | 0.457 | 0.465 |
| zuzüglich | 4210 | 892 | 0.212 | 0.187 | 0.127 | 0.148 |

Table 1: Comparison of model fit (Nagelkerke's pseudo $R^2$) between 3 different models for each preposition: a) a model with 60 COReX predictors; b) a model with 7 FA predictors; and c) a model with 7 FA predictors plus 53 predictors containing random data (noise).

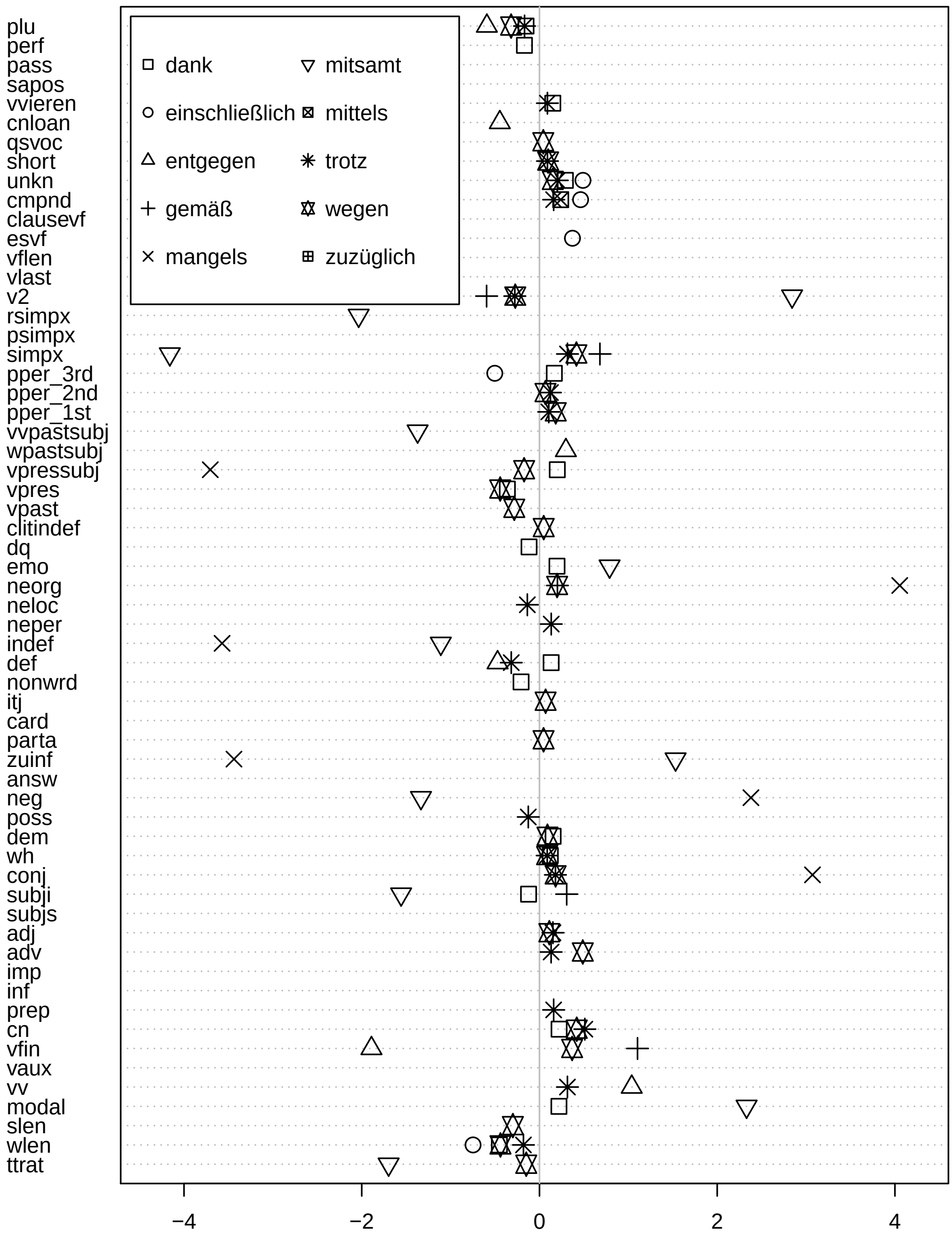


Figure 5: COReX features: coefficient estimates with associated p-value < 0.05; a separate model was specified for each preposition.

### 4.2 Measure nouns

This case study represents a replication of **Schaefer2018**. The authors models case alternation in German measure noun phrases (MNPs) based on data from the DECOW corpus (Schäfer & Bildhauer 2012). Only MNPs with a kind-denoting noun (such as *Wein* 'wine') and a measure noun (such as *Glas* 'glass') similar to English *a glass of wine* are considered. In the first alternant, the kind noun is in the genitive as in (7a). In the second alternant, the kind noun shows case identity with the head measure noun as in (7b).[9]

(7) a. Wir trinken [[ein Glas] [guten Weins]].
we drink a glass good wine
We drink a glass of good wine.

b. Wir trinken [[ein Glas] [guten Wein]].

It is noteworthy that the alternation only occurs if the embedded kind noun is determinerless but is modified by an adjective (**Schaefer2018**). Based on theoretical analysis, the author names a number of influencing factors for this alternation such as whether the whole MNP has a cardinal determiner, which case it stands in as a whole, and how strongly the individual lexical items favour a specific variant. Based on previously published assumptions (e. g., Hentschel 1993, Zimmer 2015), he also models the effect of register or style. It is generally assumed that certain registers or styles favour the genitive alternant over the alternant with case identity. However, given the annotation available in the corpus, only two weak indicator variables are used as proxies, namely `Badness` (a measure of document quality available for all DECOW documents; see **SchaeferEa2013**) and `Genitives` (a measure of the frequency of genitives). Only *Genitives* passes the tests in the original study, and it is a weak effect (**Schaefer2018**). These indicator variables are effectively like a highly reduced COReX feature set, and we attempt replicate the study here in our exploration of the usefulness of the COReX features.

#### 4.2.1 Comparing aggregated with non-aggregated predictors

For this case study, we selected all combinations of measure nouns and kind-denoting nouns that occurred at least 10 times in the published dataset[10] of **Schaefer2018**, resulting in 107 distinct combinations such as *Haufen - Schrott* ('pile - junk') and *Schüssel - Salat* ('bowl - salad'). For these combinations, we extracted all instances of both alternants (i. e., the alternant with case identity and the genitive alternant, as illustrated in (7) above) from both the DECOW16 web corpus and DeReKo[11]. As in the case study on prepositions, we then discarded all but one instance per document in order to avoid correlated data points at the document level. This procedure yielded a total of 7906 data points, instantiated by 31 distinct measure nouns and 56 distinct kind-denoting nouns. Table 2 gives the distribution of construction type by corpus. The proportion of the genitive alternant in DECOW16B is considerably lower than in DeReKo. Given the plausible assumption that the two corpora differ markedly with respect to the mixture of registers contained in them (with DeReKo leaning more towards the kind of formal, well-redacted language typical of newspaper articles), the observed distribution of alternants lends further support to the idea that the alternation is, at least to some degree, influenced by register and/or style.

Again, we model the probability of a particular alternative (the genitive alternant, in this case) given different kinds of document-level information. As before, we specify a model using the seven document scores obtained from the factor analysis as predictors, and in a separate model, we directly use the 60 document-level counts obtained from COReX. In addition, we

[9] Examples are taken from **Schaefer2018**.
[10] DOI/URL: https://dx.doi.org/10.5281/zenodo.1254871
[11] Again, we used the subset of **Bubenhofer-ea2014**

| | Corpus | |
|---|---|---|
| Construction | DECOW16B | DeReKo |
| Case identity | 4360 | 1394 |
| Genitive | 1272 | 880 |

Table 2: Distribution of alternants by corpus. The proportion of the genitive alternant in DeReKo (0.39) is considerably higher than in DECOW16B (0.23).

| Total | Genitive | p(Genitive) | $R^2_{COReX}$ | $R^2_{FA}$ | $R^2_{FA+Noise}$ |
|---|---|---|---|---|---|
| 7906 | 1737 | 0.27 | 0.214 | 0.115 | 0.125 |

Table 3: Comparison of model fit (Nagelkerke's pseudo $R^2$) between three different models for each preposition: a) a model with 60 COReX predictors; b) a model with 7 FA predictors; and c) a model with the same 7 FA predictors plus 53 predictors containing random data from a normal distribution.

specify a third model in which another 53 predictor variables are added to the 7 predictors of the FA-based model. Table 3 shows the comparison of model fit.

$$P(Genitive = 1) = logit^{-1}(\alpha + \beta_1 c_1 + \beta_2 c_2 + \ldots + \beta_{60} c_{60}) \tag{8}$$

Matrizen oder Laufindex oder Fußnote und erklären

$$P(Genitive = 1) = logit^{-1}(\alpha + \beta_1 f_1 + \beta_2 f_2 + \ldots + \beta_7 f_7) \tag{9}$$

$$P(Genitive = 1) = logit^{-1}(\alpha + \beta_1 f_1 + \beta_2 f_2 + \ldots + \beta_7 f_7 + r_1 \beta_8 + r_2 \beta 9 \ldots + r_{53} \beta_{46}) \tag{10}$$

## References


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

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.

Hentschel, Elke. 1993. Flexionsverfall im Deutschen? Die Kasusmarkierung bei partitiven Genitiv-Attributen. *Zeitschrift für Germanistische Linguistik* 21(3). 320–333.

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

Müller, Stefan. 2010. *Grammatiktheorie*. Staufenburg. `http://hpsg.fu-berlin.de/~stefan/Pub/grammatiktheorie.html`.

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.

Schäfer, Roland & Felix Bildhauer. 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 & Stelios Piperidis (eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 486–493. Istanbul: ELRA.

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.

Zimmer, Christian. 2015. Bei einem Glas guten Wein(es): Der Abbau des partitiven Genitivs und seine Reflexe im Gegenwartsdeutschen. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 137(1). 1–41.