Roland Schäfer*

# Competing Constructions for German Measure Noun Phrases: from Usage Data to Experimental Validation

**Abstract:**

**Keywords:** corpus methods and experimental methods, self-paced reading, forced choice, alternations, hierarchical models, measure constructions, German

# 1 Cognitively oriented corpus linguistics

## 1.1 Corpora and cognition

This paper deals with a morpho-syntactic alternation that occurs only in a very specific syntactic measure noun phrase construction in German. By *alternation* I refer to a situation where two or more forms or constructions are available with no clear difference in acceptability, function, or meaning.

– usage-based/cognitive corpus linguistics and the study of alternations
– validationist turn; 2016 special issue CoGL, but Bresnan did it all the time!
– also Gries (2015b)
– the value of case studies
– ‘Zweifelsfälle’ (Klein, 2009) understood as ideal cases cor cognitive modeling
– why forced choice and why reading times?

## 1.2 A word on statistical analysis

In this section, I motivate the choice of statistical models which I use later in Section 3. Readers might think that the methods used here for modeling a grammatical alternation – namely Hierarchical Regression and Generalised Linear Mixed Models (GLMMs) – does not require much motivation. After all, GLMMs have been established as the major tool in the analysis of alternation phenomena (Bresnan et al., 2007; Bresnan & Hay, 2008; Bresnan & Ford, 2010;

***Corresponding author: Roland Schäfer,** Freie Universität Berlin

Divjak & Arppe, 2013; Gries, 2015a; Nesset & Janda, 2010, to name just a few publications). Over the past few years, however, modified or alternative methods have been proposed, and I want to make a few remarks on Bayesian estimation (see Gelman et al., 2014) as suggested in Levshina (2016); Divjak (2016), for example. Conceptually, I see three points of discussion that should be kept apart. First, Bayesian methods are sometimes touted as superior tools for scientific inference compared to frequentist methods. Second, it has been proposed that the Bayesian interpretation of probability is more cognitively adequate for the modeling of linguistic data. Third, and very specific to this paper, given established methods in the modeling of alternation and variation, it has to be decided whether so-called Bayesian methods lead to substantially different results.

As for the first question, this is not quite the place to discuss it fully. The basic distinction is a philosophical one and related to direct and inverse probability (e. g., Senn, 2011). Frequentists assume that models and parameters are fixed, for example a model specifiying that a coin is fair. Given we know the distribution of the values measured, we can calculate for observed data (for example a measurement of 3 heads in an experiment with 10 tosses) how often such a result or a more extreme result would occur if the model were true and we repeated the experiment arbitrarily often. This is essentially the frequentist notion of direct probability, i. e., long-run frequencies under replication. What makes this attractive is that a model – specified with reference to a substantive theory of the given phenomenon – can fail on given data, and if it does not, the outcome of the experiment contributes to the overall well-testedness of the theory. Bayesian approaches, on the other hand, condition on the data and quantify (inductively) the probability of models given specific observations. The models are thus not fixed, and the probability is usually equated with researchers posterior beliefs about models. There is actually a debate among Bayesians whether and how a notion of theory testing is compatible (or even already contained) in the Bayesian approach. In Gelman & Shalizi (2013, 10), the authors acknowledge this and state about the standard inductive interpretation of Bayesianism that *most of this received view of Bayesian inference is wrong*, and they develop a Bayesian notion of p values (see also Mayo, 2013, for a frequentist reply; also Senn, 2011 on different strands of Bayesianism and their stance on inductive vs. deductive reasoning). Clearly, in such quarrels between and among camps of philosophers of science and statisticians it is virtually impossible for mere practitioners to take sides based on a sufficiently informed judgement.

These quarrels relate to the second question, however. Divjak (2016, 301–302) speaks favourably of Bayesian methods because the Bayesian concept of probability is more adequate for cognitive modeling compared to the frequentist one. This criticism is part of a larger body of literature asking for cognitively plausible modeling techniques, for example Naive Discriminative Learning (NDL; Baayen, 2011; Baayen et al., 2013; Milin et al., 2016; Theijssen et al., 2013). Frequentist methods were, however, never designed as cognitive models (see above). The actual question that emerges here is what our statistical models (estimated on corpus data) should represent. Are they inductive models of cognitive representations that learners would infer from being exposed to the corpus data?[1] Or are they tests of theories that are pre-specified and merely tested for predictive accuracy on output data. In the former case, we should definitely resort to methods like NDL, which most likely would entail tossing most previous work done in (cognitive) linguistics into the bin and abandoning all high-level generalisations that most existing studies have been based upon. In fact, arguments to this extreme effect and against using high-level generalisation have been made, for example in Baayen et al. (2016), Divjak (2016, 299–300), Ramscar & Port (2016), Theijssen et al. (2013). In the latter case, the cognitive commitment does not necessarily extend to the statistical methods used. These methods then do not need to be any more cognitively plausible than an ANOVA used to analyse the results from an experiment. I view my own work in the tradition of testing theories (embracing high-level generalisations) on data (see Sections 2–4), but I am fundamentally agnostic with respect to this question. The best strategy for cognitive linguistics as a field might be to cultivate many methods while making sure that each method is applied carefully and competently.

The third point is the most practically relevant in the context of this paper. In her otherwise excellent study, Levshina (2016, 251–252) argues for Bayesian estimation in mixed regression settings. First, she claims that *while frequentist statistics only allows one to test whether the null hypothesis can be rejected, Bayesian statistics enables one both to test the null hypothesis and to estimate the probability of specific parameter values given the data.* This does not do justice to frequentist methods in that rejection of the null hypothesis is characteristic of Fisher's approach, but in the Neyman-Pearson approach, results ideally *favour* the main hypothesis (vs. the alternative), cf. Lehmann (1993, 2011); Perezgonzalez (2015). She then explains that a *distinctive feature of Bayesian statistics is the use of so-called priors* and that *posterior probabilities depend on both the prior beliefs and the data, whereas the results of a frequentist model depend only on the data* (Levshina, 2016, 252). Remarkably, however, she does *not* use informative priors, and in her footnote 8 (Levshina, 2016, 252) admits that pri-

**1** In which case we are doing *data science in language research* in the words of Milin et al., 2016.

ors were probed using trial and error, such that the proclaimed advantage of Bayesian modeling was apparently not taken advantage of.[2] Now, Maximum Likelihood Estimation (MLE), the traditional method which could have been used here, is not exactly *frequentist* in the sense of Neyman-Pearson test theory. MLE also conditions on the data in as much as it finds the most likely set of parameters given the data. What is more, Bayesian estimators in fact incorporate the Likelihood function and merely multiply it by the prior (Gelman et al., 2014, PAGENUMBER??). If the prior is flat, results converge (see also Gelman & Hill, 2006, 347). The same is true if the sample size is large compared to the number of parameters, at least for finite-dimensional models (Freedman, 1999, 1119–1120), a well-established result known as the Bernstein-von Mises theorem. With a modest model structure including 17 fixed effects and $2,646$ data points in Levshina (2016), it is highly likely that the same results would have obtained with Maximum Likelihood methods. If not, especially in case of convergence problems, this would have been an ideal occasion to demonstrate the superiority of Bayesian estimation. There are situations where Bayesian estimators can be more robust, namely with heavily censored data, empty cells, complex hierarchical models, perfect separation (see Freedman, 1999, (Gelman & Hill, 2006, 345–348)). I want to reiterate that these points do not in any way invalidate the study presented by Levshina (2016). However, being *Bayesian* is not among its selling points. In Section 3.3, I compare my own hierarchical models estimated with Bayesian and MLE estimators to demonstrate convergence.

This concludes the general introduction. In the remainder of the paper, I introduce the alternation phenomenon (Section 2) and suggest a theory-driven set of factors influencing the alternation (Section 2.2). Then, the corpus study is presented including statistical analysis (Section 3). The two experiments validating the corpus study are then reported (Section 4 before I sum up the paper in Section 5.

**2** In the words of Senn (2011): *You may believe you are a Bayesian but you are probably wrong.* (Gelman & Hill, 2006, 347–348) *view any noninformative prior distribution as inherently provisional.*

# 2 Case assignment in German measure NPs

## 2.1 Two stable cases and a case alternation

In this section, I introduce and illustrate the relevant alternating constructions. I describe the narrowly defined syntactic configuration in which the alternation occurs, and I motivate the focus on *only* this narrow (range rather than, for example, the whole range of nominal constructions expressing quantities).

I use the term *measure noun phrase* (MNP) to refer a noun phrase (NP) in which a kind-denoting (count or mass) noun depends on another noun that specifies a quantity of the objects or the substance denoted by the kind noun. I call the kind-denoting noun simply the *kind noun* and the quantity-denoting noun the *measure noun*. Measure nouns can be all sorts of nouns which genuinely denote a quantity (such as *litre* or *amount*) but also nouns denoting containers, collections, etc. (such as *glass* or *bucket*). In a similar vein, Brems (2003, 284) considers nouns as measure nouns *which, strictly speaking, do not designate a 'measure', but display a more nebulous* (*sic!*) *potential for quantification* (see also Koptjevskaja-Tamm, 2001, 530, and Rutkowski, 2007, 338). For illustration purposes, in the English *a cup of fine coffee*, *cup* is the measure noun, and *coffee* is the kind noun.

In the case at hand, three different syntactic configurations need to be distinguished w. r. t. case assignment inside German measure noun phrases. If the kind noun forms an NP with a determiner, the construction resembles (and is usually called) a *pseudo-partitive* (on partitives and pseudo-partitives see, e. g., Barker, 1998; Selkirk, 1977; Stickney, 2007; Vos, 1999; for a recent application of the terminology to German, see Grestenberger, 2015).[3] Here, the kind noun is in the genitive, and I refer to the construction in (1a) as the *Pseudo-partitive Genitive Construction* (PGC).

(1) a. Wir trinken [[eine Tasse]$_{\text{Acc}}$ [eines leckeren Kaffees]$_{\text{Gen}}$]$_{\text{Acc}}$.
we drink a cup a tasty coffee
We drink a cup of a tasty coffee.

3 If the kind noun is definite, the construction instantiates a true partitive. Whereas partitives are constructions denoting a proper part-of relation as in *a sip of the wine*, pseudo partitives – albeit syntactically similar and diachronically related to partitives in many languages – merely denote quantities and contain indefinite kind nouns as in *a sip of wine*.In the literature on German, some authors incorrectly call the pseudo-partitive a *partitive* Hentschel (1993) while some realise the difference and at least mention it (Eschenbach, 1994; Gallmann & Lindauer, 1994; Löbel, 1989; Zimmer, 2015).

b. * Wir trinken [[eine Tasse]$_{\text{Acc}}$ [einen leckeren Kaffee]$_{\text{Acc}}$]$_{\text{Acc}}$.

If the kind noun is bare – i. e., if it does neither come with a determiner nor a modifying adjective – it has to agree in case with the measure noun, and the genitive seen in the PGC is not acceptable, see (2).

(2) a. * Wir trinken [[eine Tasse]$_{\text{Acc}}$ [Kaffees]$_{\text{Gen}}$]$_{\text{Acc}}$.
b. Wir trinken [[eine Tasse]$_{\text{Acc}}$ [Kaffee]$_{\text{Acc}}$]$_{\text{Acc}}$.
we drink a cup coffee
We drink a cup of coffee.

This construction is usually classified as a *Narrow Apposition Construction* (Löbel, 1986), henceforth NAC.[4] Notice that the unavailability of the genitive on the kind noun can be seen as following from a rather quirky constraint that genitive NPs in German require the presence of some strongly case-marked element (determiner or adjective) in addition to the head noun in order to be acceptable (Gallmann & Lindauer, 1994; Schachtl, 1989). While only an accusative is shown in (2), the kind noun obligatory agrees in case also with nominative and dative measure nouns.

The actual alternation can be observed *only when the kind noun occurs with an attributive adjective but without a determiner*, as in (3), where both the NAC in (3a) and the PGC in (3b) are equally acceptable. They are by-and-large functionally and semantically equivalent. However, Section 2.2 is devoted to developing hypotheses about subtle differences between them.[5]

(3) a. Wir trinken [[eine Tasse]$_{\text{Acc}}$ [heißen Kaffee]$_{\text{Acc}}$]$_{\text{Acc}}$.
we drink a cup hot coffee
We drink a cup of hot coffee.
b. Wir trinken [[eine Tasse]$_{\text{Acc}}$ [heißen Kaffees]$_{\text{Gen}}$]$_{\text{Acc}}$.

**4** The construction as in (2b) is also referred to as the *Direct Partitive Construction* for other Germanic languages in which the PGC with the synthetic genitive is not available. This nomenclature makes sense in contrast to the *Indirect Partitive Construction* with prepositional linkers translating to 'of' – i. e., analytic genitives – in such languages, see Hankamer & Mikkelsen (2008) for Danish. For German, this terminology is not distinctive enough, which is why I use the terms NAC and PGC.

**5** Some descriptive and normative grammars take stronger positions w. r. t. the acceptability of the two options. See Hentschel (1993); Zimmer (2015) for analyses of the sometimes absurd stances taken in grammars of German. As the usage and experimental data presented below (especially Section 4.1) should render it clear, there might be preferences under certain circumstances, but we cannot assume either construction to be unacceptable.

| kind NP is: | bare noun NP [...$N_{meas}$ [$N_{kind}$]] | NP with adjective [...$N_{meas}$ [AP $N_{kind}$]] | NP with determiner [...$N_{meas}$ [D $N_{kind}$]] |
|---|---|---|---|
| narrow apposition | $NAC_{bare}$ | $NAC_{adj}$ | — |
| pseudo-partitive genitive | — | $PGC_{adj}$ | $PGC_{det}$ |

**Table 1:** Distribution of the NAC and PGC constructions in different NP structures

The distribution of the case patterns in the NAC (case identity between the measure noun and the kind noun) and in the PGC (genitive on the kind noun, regardless of the case of the measure noun) depending on the structure of the kind NP are summarised in Table 1. I call the narrow apposition construction with a bare kind noun the $NAC_{bare}$, the partitive genitive with a determiner in the kind noun phrase the $PGC_{det}$. I call the alternating variants with an adjective but no determiner in the kind noun phrase $NAC_{adj}$, and $PGC_{adj}$, respectively. This paper is about the middle column of Table 1, i. e., the syntactic configuration in which two different case patterns are acceptable.

I now turn to some more subtle issues related to the measure noun case alternation, namely:

i. alleged alternatives to the $NAC_{adj}$ and the $PGC_{adj}$
ii. alternative *weak* forms of dative singular neuter adjectives
iii. similar constructions with plural/collective kind nouns
iv. grammaticalised non-inflected measure nouns
v. alternative constructions for expressing quantities

First of all, (i) refers to claims found in some grammars that a generic nominative, accusative, and even dative on the kind noun are used instead of the genitive ($PGC_{adj}$) or case agreement ($NAC_{adj}$). Overviews can be found in Hentschel, 1993; Zimmer, 2015. It was shown empirically in Hentschel (1993) that such variants are de facto not acceptable. Also, in my corpus sample, they simply did not occur. Even if they are accepted by some speakers, their extremely low frequency makes it virtually impossible to study them using either corpus methods (Section 3) or experimental approaches (Section 4), and I consequently ignore them.

As for (ii), a complication with neuter kind nouns in the dative is mentioned by Zimmer (2015, 20–22). In the $NAC_{adj}$, the adjective normally inflects with the so-called *strong* case and number suffixes, which are used when no determiner with strong inflection is present in the NP. Zimmer (2015) reports a high number of occurrences of adjectives being inflected with the *weak* inflectional marker, which are normally only used if a strongly inflected determiner precedes the adjective: *kalt-en Wasser* ‘cold water’ instead of *kalt-em Wasser*.[6] In my corpus sample, this tendency was not nearly as clear, and there was a high number of very noisy sentences among those potentially showing this pattern. Hence, I do not discuss these forms.[7]

Turning to (iii), readers might have noticed that so far only MNPs denoting quantities of substances (mass kind nouns such as *ein Glas roter Wein/roten Weines* ‘a glass of red wine’) have been discussed. If the kind noun is a plural count noun as in *ein Sack kleine Äpfel* ($\text{NAC}_{\text{adj}}$) or *ein Sack kleiner Äpfel* ($\text{PGC}_{\text{adj}}$) ‘a bag of small apples’, a similar alternation between PGC and NAC can be observed. In line with experimental results reported in Zimmer (2015, 15–16), I found that the PGC is so dominant with plural kind nouns (67 of 861 cases or over 92%, cf. Section 3) that the alternation cannot be analysed in the same way as in the singular case.[8] While this will play a role in the interpretation of the corpus findings, MNPs with plural kind nouns will not be included in the corpus study and the experiments reported here.

As for (iv), some measure nouns have been grammaticalised in a way that they always appear non-inflected. They are typical measure nouns like *Gramm* ‘gram’, *Pfund* ‘pound’ or *Prozent* ‘percent’, which do not have plural forms at all. I treat these cases like other measure nouns because they enter into both the $\text{NAC}_{\text{adj}}$ as in (4a) and the $\text{PGC}_{\text{adj}}$ as in (4b). In Section 2.2, degrees of grammaticalisation as a factor influencing the alternation will be discussed, however.

(4) a. zwei Gramm brauner Zucker
   b. zwei Gramm braunen Zuckers
      two gram brown sugar
      two grams of brown sugar

Finally, (v) suggests that there are alternative ways of expressing similar quantificational meanings. In the variationist tradition, which is strongly influenced by Labovian sociolinguistics, the *principle of accountability* would dictate that proper studies should examine a variationist *variable*, i. e., all different *ways of saying the same thing* (Labov, 1966, Labov, 1969, for an overview see Taglia-

**6** See Section 2.2 for more discussion of adjectival inflection.

**7** All corpus data and scripts used for this paper will be released freely, so further examination of the few cases maybe showing this kind of inflection is possible.

**8** More concretely, I these cases were included in the regression analysis of the corpus data along with a factor encoding the number of the kind noun, this factor would most assuredly override any other regressor for the data points with a plural kind noun.

monte, 2012). In the case at hand, the variable might be something like *measuring quantities of substances and collections*. I argue that it is fully justified to focus narrowly on $\text{NAC}_{\text{adj}}$ and $\text{PGC}_{\text{adj}}$ with their well-defined morpho-syntactic properties, mostly because the alternative constructions are not used in the same range of contexts. Two major alternatives might be considered. First of all, the analytic (pseudo-)partitive with *von* 'of' is only available as an alternative to the PGC if the kind noun phrase contains a (definite or indefinite) determiner as in (5).

(5) a. ein Glas von dem roten Wein
a glass of the red wine
a glass of the red wine

b. * ein Glas von rotem Wein
a glass of red wine
a glass of red wine

This means that the *von* (pseudo-)partitive does not compete with the $\text{NAC}_{\text{adj}}$ and the $\text{PGC}_{\text{adj}}$. In fact, there is not a single context in which they can be interchanged.

Second constructions with *voll* (*von*)/*voller* 'full (of)' and *mit* 'with' are available, see (6).

(6) a. ein Glas voll von rotem Wein
a glass full of red wine
a glass full of red wine

b. ein Glas voll/voller rotem/roter Wein
a glass full-of red wine
a glass full of red wine

c. ein Glas mit rotem Wein
a glass with red wine
a glass with red wine in it/a glass filled with red wine

These construction have very idiosyncratic properties (the construction with *voller* is discussed in Zeldes, to appear, the construction with *mit* is discussed in Bhatt, 1990), they are not semantically equivalent to the NAC and the PGC, and they are only available for a small subset of measure nouns. All of these constructions are only used with measure nouns denoting containers, and the whole NP always refers to the container, and never to the quantity contained in it. They are consequently incompatible with measure nouns denoting natural portions such as *Schluck* 'gulp' or *Haufen* 'heap' and strongly grammaticalised nouns such as *Gramm* 'gram'. This disqualifies them as alternatives to the $\text{NAC}_{\text{adj}}$ or

$\mathrm{PGC_{adj}}$ in most contexts, and they are thus decidedly *not* more ways of saying the same thing. It is therefore reasonable to focus on the two well-defined variants.

This concludes the descriptive overview of the phenomenon. I have shown that there is an alternation between two measure noun constructions in a narrow syntactic configuration (kind NP with an adjective but without a determiner), and that the two constructions differ in the case of the kind noun (case agreement with the measure noun or genitive). I turn to some more theory-oriented discussion in the next section.

## 2.2 Factors controlling the alternation

This section briefly reviews existing analyses of the $\mathrm{PGC_{adj}}$vs. $\mathrm{NAC_{adj}}$ alternation and related issues. I also develop my own analysis and the appropriate hypotheses for the empirical studies presented in Sections 3 and 4. I discuss four main aspects: first, the different syntactic structures of the variants $\mathrm{NAC_{adj}}$ and $\mathrm{PGC_{adj}}$ in relation to their syntactic prototypes $\mathrm{NAC_{bare}}$ and $\mathrm{PGC_{det}}$; second, degrees of grammaticalisation of the measure noun associated with the two prototypes; third, a semantic prototypicality effect of the degrees of grammaticalisation (preference to occur with cardinals); fourth, preferences in different registers.

My whole analysis is based on the idea that the syntactic structure $[\mathrm{N}_1\ [\mathrm{A}\ \mathrm{N}_2]_{\mathrm{NP}_2}]_{\mathrm{NP}_1}$ as instantiated in the $\mathrm{NAC_{adj}}$ and the $\mathrm{PGC_{adj}}$ is ambiguous. To show this, the first question is which noun constitutes the head of the whole construction. This was answered quite clearly by Löbel (1986) already (see also Eschenbach, 1994, 213, and Gallmann & Lindauer, 1994, 16). Subject-verb agreement is always realised on the measure noun, and we can therefore assume that the measure noun is the head. However, the MNP-internal structure has more interesting cues to offer if the strong/weak inflection patterns of adjectives are taken into account. In NPs with a strongly inflected determiner, attributive adjectives inflect according to the massively syncretistic *weak* pattern. If there is no determiner (as is the case in the alternating constructions), attributive adjectives inflect like determiners themselves, however. This is called the *strong* inflectional pattern. Thus, the adjectives in the $\mathrm{NAC_{adj}}$ and the $\mathrm{PGC_{adj}}$ have properties of adjectives as well as determiners. On the one hand, they are lexical adjectives and function as attributive modifiers. On the other hand, they are inflected like determiners, and they are the leftmost element in the NP, which is typical of determiners. This unusual double nature of adjectives in NPs without determiners leads to a plausible probabilistic interpretation of the pattern shown in Table 1. Whenever speakers classify the adjective in the kind noun phrase more as a determiner, they have to use the $\mathrm{PGC}_{\mathrm{adj}}$ because, if there is a determiner, the PGC (just like in the case of the $\mathrm{PGC}_{\mathrm{det}}$) is the only option. When they classify the adjective more as an adjective, the kind NP has no determiner, and they have to use the $\mathrm{NAC}_{\mathrm{adj}}$ (like in the case of the $\mathrm{NAC}_{\mathrm{bare}}$).[9]

This morpho-syntactic ambiguity means that the $\mathrm{NAC}_{\mathrm{adj}}$ is in fact a $\mathrm{NAC}_{\mathrm{bare}}$in disguise, and the $\mathrm{PGC}_{\mathrm{adj}}$ is a $\mathrm{PGC}_{\mathrm{det}}$in disguise. This should ideally be reflected in the following basic selection effect: It is known from frameworks like Collexeme Analysis (Gries & Stefanowitsch, 2004) that lemmas are attracted with different strengths by competing constructions. If the $\mathrm{NAC}_{\mathrm{adj}}$ is highly similar to the $\mathrm{NAC}_{\mathrm{bare}}$ and the $\mathrm{PGC}_{\mathrm{adj}}$ is highly similar to the $\mathrm{PGC}_{\mathrm{det}}$, the probability with which individual noun lemmas appear in the alternating constructions should be predictable at least partly from the relative frequency with which the same lemmas are used in the non-alternating constructions (see Levshina, 2016, 246–249 for a similar idea in a different context). Nouns which occur proportionally more often in the $\mathrm{PGC}_{\mathrm{det}}$ should favour the $\mathrm{PGC}_{\mathrm{adj}}$, and nouns which occur proportionally more often in the $\mathrm{NAC}_{\mathrm{bare}}$ should favour the $\mathrm{NAC}_{\mathrm{adj}}$. This basic attraction effect will be quantified in the corpus study Section 3.

In a probabilistic framework like Prototype Theory, the crucial question (beyond simple lemma preference effects) is, however, what controls speakers' decisions to use either variant. In the remainder of this section, I argue that the NAC and the PGC are prototypes associated with different degrees of the grammaticalisation of the measure noun, a related morpho-syntactic property, and register effects. The degree of similarity of a given instance to either of the two prototypes makes speakers chose the $\mathrm{NAC}_{\mathrm{adj}}$ or the $\mathrm{PGC}_{\mathrm{adj}}$. I first turn to grammaticalisation. It is often assumed that pseudo-partitives arise as a form of grammaticalised partitives (e. g., Koptjevskaja-Tamm, 2001, 536–539 for Finnish and Estonian, Koptjevskaja-Tamm, 2001, 559 for European languages per se). The alternating constructions in German are both clearly pseudo-partitives, but the grammaticalisation paths uncovered by (Koptjevskaja-Tamm, 2001, esp. 526–530) are still relevant in the case at hand. The grammaticalisation path can start out (in some languages) with constructions involving two referential nouns (not necessarily forming a single and contiguous NP) and a *separative* meaning

**9** While the generative analysis presented in Bhatt (1990) cannot properly deal with probabilistic effects, Bhatt comes close to this interpretation by analysing the kind NP in the GPC as a DP and in the NAC as an NP.

as in *(cut) two slices from the cake* (Koptjevskaja-Tamm, 2001, 535). The *part-of* meaning of true partitives as in *a slice of the cake* represents the first stage of a development wherein the measure noun already tends to loose some semantic content. The pseudo-partitive stage finally instantiates a *quantity-of* relation, potentially even leading to fully grammaticalised quantifiers such as *a lot.* In German, the PGC is clearly the older variant (Zimmer, 2015). It still has the potential to form a true partitive (if the kind noun is definite). Conversely, the NAC completely lacks this ability to form true partitives, and it forms a more restrictive environment with less combinatory and semantic flexibility. Hence, we can expect the $\mathrm{NAC}_{\mathrm{adj}}$ to be a prototypical hosting construction for more strongly grammaticalised measure nouns. For example, highly grammaticalised non-referential nouns like *Gramm* 'gram' and *Meter* 'metre' should occur proportionally more often in the $\mathrm{NAC}_{\mathrm{adj}}$ than in the $\mathrm{PGC}_{\mathrm{adj}}$.

Furthermore, the grammaticalisation path as described above leads from NPs denoting individuated objects standing in a *part-of* relation to a construction with a more diffuse *quantity-of* relation. Both types of relations can be numerically quantified – in as much as a precise number of *parts* or a numerically exact *quantity* can be specified –, but it is much more prototypical of quantities to be specified with numerical precision. Since the $\mathrm{NAC}_{\mathrm{adj}}$ is more closely associated with the *quantity-of* relation, cardinals as attributes of the measure noun are expected to have a higher proportional frequency in the $\mathrm{NAC}_{\mathrm{adj}}$. For illustration, (7) shows the expected variants under this hypothesis. In (7a), the measure noun is modified by a cardinal *drei* 'three', and hence the $\mathrm{NAC}_{\mathrm{adj}}$ is preferred. In (7b), the measure noun is modified by a non-cardinal determiner *einige* 'some', and the $\mathrm{PGC}_{\mathrm{adj}}$ is preferred.

(7) a. [[Drei Löffel]$_{\mathrm{Nom}}$ [heißer Rum]$_{\mathrm{Nom}}$]$_{\mathrm{Nom}}$ sind genug.
three spoons hot rum are enough.
Three spoonful of hot rum are enough.

b. [[Einige Löffel]$_{\mathrm{Nom}}$ [heißen Rums]$_{\mathrm{Gen}}$]$_{\mathrm{Nom}}$ sind genug
some spoons hot rum are enough.
A few spoonful of hot rum are enough.

The statistical models presented in Section 3 will be specified in a way that they could detect such a preference.

Finally, it has often been iterated that the $\mathrm{PGC}_{\mathrm{adj}}$ is more typical of higher registers or even exclusive to written language (see Hentschel, 1993, 320–323). This is not surprising in as much as the genitive – an intrinsic part of the PGC – is generally underrepresented in colloquial vernacular variants of German as a result of a diachronic process wherein many (but not all) uses of the genitive are replaced by other cases or periphrastic constructions (Fleischer & Schallert, 2011). Under an integral view of prototypes, which incorporate effects related to larger contexts and registers, such preferences can be to be part of the construction prototypes. In the corpus study in Section 3, register effects will therefore be modeled – even if only using two very simple proxy variables.

To summarise the discussion and core hypotheses, I assume that the alternation is controlled by the similarity of the chosen lemmas (including the degree of grammaticalisation of measure nouns), certain morpho-syntactic choices, as well as the larger utterance context to two prototypes instantiated more straightforwardly by the two non-alternating cases of NAC and PGC. Concretely, I predict that:

i. The relative frequencies with which measure noun lemmas and kind noun lemmas appear in the prototypical (non-alternating) $\text{PGC}_{\text{det}}$ and $\text{NAC}_{\text{bare}}$ are predictive of the probability with which the alternating $\text{PGC}_{\text{adj}}$ and the $\text{NAC}_{\text{adj}}$ are chosen.
ii. (Classes of) more strongly grammaticalised measure nouns favour the $\text{NAC}_{\text{adj}}$.
iii. Measure nouns modified by cardinals favour the $\text{NAC}_{\text{adj}}$.
iv. The $\text{NAC}_{\text{adj}}$ is associated with higher registers.

In the next section, I report the corpus study that was designed to test these hypotheses on usage data. The resulting models will then be validated using experimental methods in Section 4.

# 3 Corpus study

## 3.1 Corpus choice and sampling

For the present study, I used the German *Corpus from the Web* (COW) in its 2014 version DECOW14A (Schäfer & Bildhauer, 2012; Schäfer, 2015 and Biemann et al., 2013; Schäfer & Bildhauer, 2013 for overviews of web corpora in general and the methodology of their construction), which contains almost 21 billion tokens.[10] I chose this corpus for two main reasons.[11] First, the ex-

**10** The corpora are made available for free at https://www.webcorpora.org. At the time of this writing, a newer 2016 version DECOW16 had already been released.

**11** The use of web data for linguistic research does require explicit and careful justification. Due to the noisy nature and unknown composition of the web, only carefully designed and ternal validity of any study is increased through a higher heterogeneity of the sample (Maxwell & Delaney, 2004, 30), and the DECOW corpus has clearly a much more heterogeneous composition compared to the only other very large corpus of German, the DeReKo (Kupietz et al., 2010) of the Institute for the German Language (IDS), which contains almost exclusively newspaper texts.[12] Second, it was already mentioned that normative grammars often adopt clear positions regarding the grammaticality of either the $\mathrm{NAC}_{\mathrm{adj}}$ or the $\mathrm{PGC}_{\mathrm{adj}}$. Thus, newspaper text or any other text that conforms strongly to normative grammars might not represent the alternation phenomenon fully (and without bias) because authors and proof-readers might favour one alternative or the other. Web corpora, on the other hand, contain at least some amount of non-standard language from forums and similar sources. For these or similar reasons, COW corpora have been used in a number of peer-reviewed publications, for example Goethem & Hiligsmann (2014); Goethem & Hüning (2015); Müller (2014); Schäfer (2016 aop); Schäfer & Sayatz (2014, 2016); Zimmer (2015). Therefore, DECOW can be considered the obvious choice for this study.

I now turn to the sampling procedure applied to obtain concordances for manual annotation and statistical analysis. Among the factors potentially influencing the alternation (see Section 2) were lemma-specific preference effects. Therefore, it was highly desirable to obtain a sample in which most of the highly frequent actually occurring combinations of kind nouns and measure nouns were represented. I applied a three-stage bootstrap process in order to obtain such a sample. It consisted of three steps:

i. bootstrapping a list of the one hundred most frequent mass nouns,
ii. bootstrapping a list of all measure nouns with which the mass nouns co-occur in the $\mathrm{NAC}_{\mathrm{bare}}$
iii. sampling the target constructions by querying each combination of mass noun and measure noun found in step (ii).

In step (i), I exported a list of all nouns in the DECOW14A01 sub-corpus sorted by their token frequency and manually went through it from the most frequent

established web corpora like the COW corpora or the SketchEngine corpora (Kilgarriff et al., 2014) should be used. Clearly, using search engine results is *bad science* for many reasons, most prominently total irreproducibility of results, as Kilgarriff (2006) pointed out more than ten years ago. Careless use of search engine results is still found, however, see De Clerck & Brems (2016, 171–175).

**12** It was shown in Bildhauer & Schäfer (2016) that, for example, the spread of topics is much smaller in DeReKo compared to DECOW.

noun downwards, selecting the first one hundred mass nouns that occurred in the list.[13] Abstract nouns which partially behave like mass nouns (like *Spaß* 'fun' or *Gefahr* 'danger') were excluded because they are usually not quantified in the same way as concrete mass nouns. The hundredth selected mass noun was *Schmuck* 'jewelry', which is the 3,054th most frequent noun in the original frequency list.

This resulting list of mass nouns was used in step (ii) to bootstrap a list of measure nouns co-occurring with the mass nouns. In order to generate this list, I utilised the fact that a direct sequence of two nouns almost always instantiates the bare-noun NAC if the second noun is a mass noun. Hence, I searched for all sequences $N_1N_2$ where $N_2$ was one of the mass noun lemmas extracted in step (i). Then, the resulting 100 lists of noun-noun combinations were each sorted by frequency in descending order and sieved manually to remove erroneous hits. From each of the 100 lists, I also removed noun-noun combinations that had a frequency below 2, except if the individual list would have otherwise been shorter than 20 noun-noun combinations. The result was a list of the most frequent 2,365 individual combinations of a measure noun and a mass noun.

In step (iii), each of these 2,365 noun–noun combinations was queried in the target constructions ($\text{PGC}_{\text{adj}}$ and $\text{NAC}_{\text{adj}}$) individually in each of the first ten slices of DECOW (roughly 10 billion tokens). In order to reduce the sample size for the manual annotation process, the final concordance was sampled from the results of these 2,365 queries. Since the mass nouns in the sample were distributed according to the usual power law, I used all hits for mass nouns occurring less than one hundred times, but randomly sampled one hundred hits for each mass noun that occurred one hundred or more times. The final sample contained 6,843 sentences, which was reduced to 5,063 in the manual annotation process due to removal of noisy material, erroneous hits and uninformative cases where the measure noun was in the genitive, in which case the $\text{NAC}_{\text{adj}}$ cannot be distinguished from the $\text{PGC}_{\text{adj}}$. Given the careful bootstrapping and sampling procedure described in this section, we can be highly sure that it contains all relevant and reasonably frequent noun–noun combinations in the target constructions.[14]

---

**13** DECOW14A01 is the first slice (roughly a twentieth) of the complete DECOW14A corpus. It contains just over one billion tokens.

**14** In a similar fashion, the 100 most frequent measure nouns occurring with plural kind nouns were bootstrapped and queried, resulting in a sample of 871 sentences. As stated in Section 2, the $\text{NAC}_{\text{adj}}$ is virtually never used with plural kind nouns, and this sample was not used except for quantifying the frequency of occurrence of the constructions (67

| Unit of reference | Variable | Type | Levels (for factors only) |
|---|---|---|---|
| Document | Badness | numeric | |
| | Genitives | numeric | |
| Sentence | Cardinal | factor | Yes, No |
| | **Construction (response)** | factor | NACa, PGCa |
| | Measurecase | factor | Nom, Acc, Dat |
| Kind lemma | Kindattraction | numeric | |
| | Kindfreq | numeric | |
| Measure lemma | Measureattraction | numeric | |
| | Measureclass | factor | Physical, Container, Amount, Portion, Rest |
| | Measurefreq | numeric | |

**Table 2:** Annotated variables for the main sample

Finally, two auxiliary samples were also drawn. As mentioned in Section 2.2, the distribution of the measure noun and kind noun lemmas in the $\mathrm{NAC}_{\mathrm{bare}}$ and the $\mathrm{PGC}_{\mathrm{det}}$ with a determiner will be modeled as factors influencing the alternation. Therefore, all noun-noun pairs from the bootstrap process were also queried in the two non-alternating constructions, resulting in 17,252 hits for the $\mathrm{PGC}_{\mathrm{det}}$ and 315,635 hits for the $\mathrm{NAC}_{\mathrm{bare}}$.

## 3.2 Variables and annotation

The full set of manually annotated variables for the main sample is given in Table 2, and I briefly discuss it now.[15] Notice that *Construction* is the response variable (or 'dependent variable') with the values *PGCa* and *NACa*.

The variables *Kindattraction* and *Measureattraction* encode the ratio with which a given kind noun lemma or measure noun lemma occurs in the $\mathrm{NAC}_{\mathrm{bare}}$ and the $\mathrm{PGC}_{\mathrm{det}}$. They were calculated from the auxiliary samples described at the end of Section 3.1 as a log-transformed quotient. The higher the value, the more often the noun occurs in the $\mathrm{PGC}_{\mathrm{adj}}$(proportionally).[16] *Kindfreq* and

times $\mathrm{NAC}_{\mathrm{adj}}$ and 794 times $\mathrm{PGC}_{\mathrm{adj}}$). The sample is distributed in the data package accompanying this paper, however.

**15** All numeric variables were also z-transformed (i. e., centered to the mean and rescaled such that they have a standard deviation of 1) to facilitate their interpretation in the regression models reported in the next section.

**16** One could argue that some more advanced measure of attraction strength should be used, as is done in Collostructional (or Collexeme) Analysis (Gries & Stefanowitsch, 2004).

*Measurefreq* are the logarithm-transformed frequencies per 1,000,000 words of each lemma, extracted from the frequency lists distributed by the DECOW corpus creators on their web page.

In Section 2.2, it was hypothesised that classes of measure lemmas might have different preferences for the two variants. To capture this, class information was annotated for measure lemmas. The classification was inspired by the list in (Koptjevskaja-Tamm, 2001, 530), but due to the low frequencies of many of the potential classes, a very coarse classification was finally used. With typical examples and their frequencies in the final sample, the classes are: *Physical* (abstract precisely measurable units such as *Liter* 'litre', *Meter* 'metre', *Gramm* 'gram'; $f = 1,968$), *Container* (*Eimer* 'bucket'; $f = 740$), *Amount* (*Menge* 'amount'; $f = 1,364$), *Portion* (natural portions like *Happen* 'bite' or *Krümel* 'crumb'; $f = 713$). The few lemmas that did not fit into either of these classes were labeled *Rest* ($f = 278$).

The variable *Cardinal* encodes whether the measure noun is modified by a cardinal ($f = 1,939$) or not ($f = 3,124$). The purpose of this variable is to test whether cardinals really favour the $\mathrm{NAC_{adj}}$ as hypothesised in Section 2.2.

To capture the influence of register or style mentioned in Section 2.2, two proxy variables were used. At the document level, the DECOW corpus has an annotation for *Badness*. As described in Schäfer et al. (2013), *Badness* measures how well the distribution of highly frequent short words in the document matches a pre-generated language model for German. Documents with higher Badness usually contain more incoherent language, shorter sentences, etc. If the PGC actually favours higher registers and styles, a high *Badness* should be correlated with fewer occurrences of the PGC. Documents in DECOW14 have also been annotated with a variable called *Genitives*. The higher the values of this variable, the lower the proportion of genitives among all case-bearing forms is. While more genitives are also indicative of higher registers, the use of this variable as a regressor in the present study might be considered problematic. Since the PGC

---

Two main points speak against such an approach. First, the attraction values will be used as regressors in a GLMM, and the values resulting from collostructional approaches, i. e., logarithmised Fisher p values, have a very unfavourable distribution for such a use. They cluster around $-\infty$ and 0, especially when (as in this case) cells contain 0 count values. This cannot be remedied on principled grounds, since, for example, a z-transformation is inadequate for such a strongly bimodal distribution (even if we clamp $-\infty$ to a very low numeric value). Second, their main use in collostructional analyses is to *sort* a list of collexemes and interpret the order. Their concrete numerical value thus does not have a solid cognitive interpretation. For the present study, it suffices to estimate the relative frequency with which speakers have encountered a specific lemma in the $\mathrm{NAC_{adj}}$ and the $\mathrm{PGC_{adj}}$.

contains a genitive itself, the regressor variable *Genitive* and the document-level variable *Genitives* are not fully independent. However, since the PGCs make up for only a minute fraction of all genitives, I still use *Genitives* as a regressor with the appropriate caveats.

Finally, one variable as added as nuisance variable in the context of the present study. It was reported in the literature that MNPs in the dative and with a masculine or neuter kind noun favour the $\mathrm{PGC}_{\mathrm{adj}}$ more than the corresponding nominative and accusative MNPs (Hentschel, 1993; Zimmer, 2015). As an example, *mit einem Stück frischen Brots* 'with a piece of fresh bread' ($\mathrm{PGC}_{\mathrm{adj}}$) would be preferred over *mit einem Stück frischem Brot* ($\mathrm{NAC}_{\mathrm{adj}}$). As with all the examples, native speakers of German will most likely notice that differences are subtle. To control for this effect, the case of the measure noun was manually annotated (variable *Measurecase*).

## 3.3 A hierarchical model of the measure noun alternation

In this section, I report the results of fitting a multilevel model to the data using R (R Core Team, 2014), *lme4* (Bates et al., 2015) for Maximum Likelihood Estimation, and *rstanarm* (Gabry & Goodrich, 2016) for Markov-Chain Monte Carlo estimation. The purpose is to model the influence of the regressors specified in Table 2 on the probability that either the $\mathrm{NAC}_{\mathrm{adj}}$ or the $\mathrm{PGC}_{\mathrm{adj}}$ is chosen. All regressors from Table 2 were included, and the measure lemma and the kind noun lemma were specified as varying-intercept random effects. The sample size was $n = 5,063$ with $1,1344$ cases of $\mathrm{PGC}_{\mathrm{adj}}$ and $3,929$ cases of $\mathrm{NAC}_{\mathrm{adj}}$. The results of the estimation are shown in Table 3 and in Figure 1. The regressors with the measure lemma as their unit of reference have no within-measure lemma variance, and the *glmer* function automatically estimates them as *group level predictors* (or *second-level effects*), cf. (Gelman & Hill, 2006, 265–269,302–304). The same goes for those listed with the kind lemma as their unit of reference. Given the coding of the response variable, coefficients leaning to the positive side can be interpreted as favouring the $\mathrm{PGC}_{\mathrm{adj}}$.

Standard diagnostics for the Maximum Likelihood Estimation (MLE) show that the model quality is quite high. Generalised variance inflation factors for the regressors were calculated to check for multicollinearity (Fox & Monette, 1992; Zuur et al., 2010), and none of the corrected $\mathrm{GVIF}^{1/2\mathrm{df}}$ was higher than 1.6. Nakagawa & Schielzeth's pseudo-coefficient of determination is $R^2_m = 0.409$ and $R^2_c = 0.495$ (see Gries, 2015a for a linguist-friendly introduction to these $R^2$ measures, or else Nakagawa & Schielzeth, 2013). The rate of correct predictions is 0.843, which means a proportional reduction of error of $\lambda = 0.297$. The measure

| Level | Regressor | $p_{PB}$ | Level | Coefficient | | CI low | | CI high | | CI contains 0 | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | MLE | MCMC | MLE | MCMC | MLE | MCMC | MLE | MCMC |
| First | Badness | 0.002 | | -0.152 | -0.155 | -0.247 | -0.247 | -0.061 | -0.065 | * | * |
| | Cardinal | 0.001 | No | 1.189 | 1.222 | 0.862 | 0.927 | 1.466 | 1.496 | * | * |
| | Genitives | 0.001 | | -0.693 | -0.711 | -0.768 | -0.801 | -0.592 | -0.616 | * | * |
| | Measurecase | 0.001 | Acc | 0.030 | 0.031 | -0.150 | -0.159 | 0.212 | 0.222 | | |
| | | | Dat | 0.705 | 0.729 | 0.455 | 0.465 | 0.944 | 0.995 | * | * |
| Second | Kindattraction | 0.020 | | 0.225 | 0.244 | 0.049 | 0.056 | 0.393 | 0.422 | * | * |
| (Kind) | Kindfreq | 0.095 | | 0.146 | 0.164 | -0.023 | -0.016 | 0.301 | 0.341 | | |
| | Kindgender | 0.001 | Neut | 0.021 | 0.013 | -0.367 | -0.409 | 0.392 | 0.435 | | |
| | | | Fem | 1.269 | 1.289 | 0.800 | 0.788 | 1.709 | 1.783 | * | * |
| Second | Measureattraction | 0.001 | | 0.282 | 0.299 | 0.106 | 0.102 | 0.447 | 0.515 | * | * |
| (Measure) | Measureclass | 0.001 | Container | 0.252 | 0.257 | -0.265 | -0.303 | 0.788 | 0.813 | | |
| | | | Rest | 0.421 | 0.379 | -0.209 | -0.378 | 1.063 | 1.091 | | |
| | | | Amount | 0.831 | 0.889 | 0.215 | 0.220 | 1.432 | 1.569 | * | * |
| | | | Portion | 1.217 | 1.253 | 0.675 | 0.689 | 1.684 | 1.840 | * | * |
| | Measurefreq | 0.005 | | -0.231 | -0.232 | -0.363 | -0.395 | -0.079 | -0.073 | * | * |

**Table 3:** Coefficient table comparing Maximum Likelihood Estimation (MLE, with 95% bootstrap confidence interval) and 'Bayesian' Markov-Chain Monte Carlo estimation (MCMC); the intercept (*Cardinal=Yes*, *Measurecase=Nom*, *Kindgender=Masc*, *Measureclass=Physical*; 0 for all numeric z-transformed regressors) is -3.548 (MLE) and -3.700 (MCMC)

lemma intercepts have a standard deviation of $\sigma_{\text{Measurelemma}} = 0.448$, the kind lemma intercept $\sigma_{\text{Kindlemma}} = 0.604$.

The coefficient estimates are specified in Table 3 for each regressor (or regressor level) in the columns labelled *Coefficient*. For a robust quantification of the precision of the estimation, I ran a parametric bootstrap (using the *confint.merMod* function from *lme4*) with $1,000$ replications, and using the percentile method for the calculation of the intervals. The resulting 95% bootstrap confidence intervals are reported in Table 3 in the columns labelled *CI low* and *CI high* (= upper and lower 2.5th percentiles). The column *CI contains 0* shows an asterisk for those intervals that do *not* include 0. Furthermore, for each regressor, a $p$ value was obtained by dropping the regressor from the full model, re-estimating the nested model and comparing it to the full model. Instead of

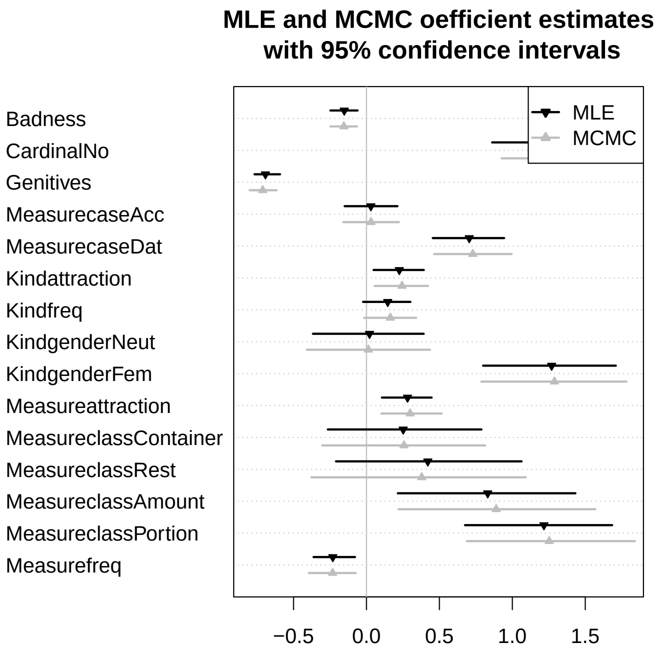


**Fig. 1:** Coefficients (MLE and MCMC) with 95% confidence intervals (for details see text); the intercept (*Cardinal=Yes*, *Measurecase=Nom*, *Kindgender=Masc*, *Measureclass=Physical*; 0 for all numeric z-transformed regressors) is -4.328 (MLE) and -4.441 (MCMC)

inexact Wald approximations or Likelihood Ratio Tests, I used a drop-in bootstrap replacement for the Likelihood Ratio Test from the function *PBmodcomp* from the *pbkrtest* package (Halekoh & Højsgaard, 2014), hence I call the corresponding value $p_{\mathrm{PB}}$. These bootstrapped $p$ values are given in the columns labelled $\mathrm{p}_{\mathrm{PB}}$ in Table 3. Only *Kindfreq* ($p_{\mathrm{PB}} = 0.095$) can be seen as slightly too high to be convincing ('non-significant').

Finally, I show that Bayesian methods in the form of Markov-Chain Monte Carlo estimation do not necessarily (and in this case predictably) lead to different results. The model was re-estimated using the *stan_glmer* function from the *rstanarm* package, which provides an *lme4*-compatible syntax for estimating common model types with the *stan* software (Carpenter et al., 2017). The algorithm was run with 4 chains and $1,000$ iterations, and I used plausible default priors. Most notably, priors for coefficients were specified as $\mathcal{N}(0,10)$ because coefficients higher than 10 or lower than $-10$ are extremely rare in well-specified models on appropriate data.[17] The algorithm converged, and for all coefficients, the $\hat{R}$ diagnostic was exactly 1. The resulting coefficients and intervals as well as an * are also given in Table 3 in the columns labelled *MCMC*. Both methods lead to exactly the same results (minus negligible numerical differences) as expected (see Section 1.2). The signs and magnitudes of the coefficients are identical, and confidence intervals have the same width and symmetry properties. Figure 1 illustrates this by also showing both estimates. Since both estimators converge, I only interpret the MLE model in the next section.

## 3.4 Interpretation

The results reported in Section 3.3 generally confirm the hypotheses from Section 2.2. First, the prototypicality effect related to the non-alternating $\mathrm{PGC}_{\mathrm{det}}$ and $\mathrm{NAC}_{\mathrm{bare}}$ can be shown, see the effect plots in Figure 2.[18] The effect is mostly as expected: if a lemma appears relatively more often in the $\mathrm{PGC}_{\mathrm{det}}$ (compared to its frequency in the $\mathrm{NAC}_{\mathrm{bare}}$), the more often the $\mathrm{PGC}_{\mathrm{adj}}$ is chosen over the $\mathrm{NAC}_{\mathrm{adj}}$ with this specific lemma. The effect for measure nouns is stronger, and it was estimated with higher precision.

**17** Consider that with a coefficient of 10, each increase by 1 in the regressor variable means an increase in odds of $exp(10) = 22,026.47$. To reliably estimate such coefficients, extremely large samples would be required.

**18** Effect plots were created using the *effects* package (Fox, 2003). They show the changes in probability for the outcome (y axis) dependent on values of a regressor (x axis), at typical values of all other regressors.

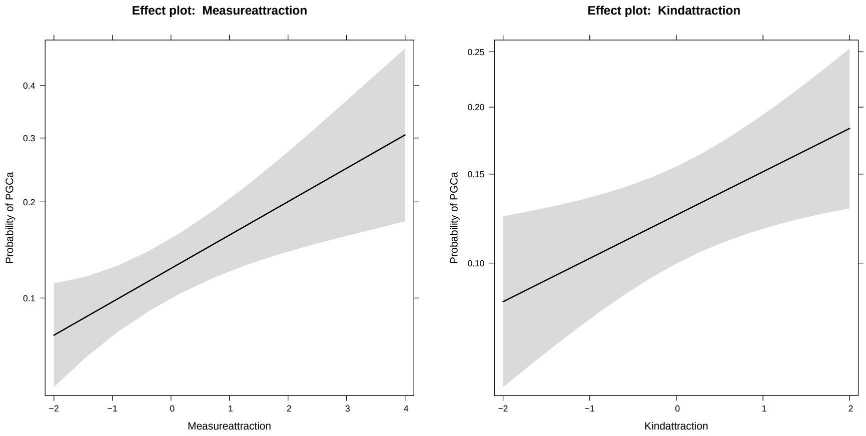


**Fig. 2:** Effect plots for the regressors *Measureattraction* and *Kindattraction*; y axes are not aligned

An interesting picture emerges for the lemma frequencies. A higher than average lemma frequency of measure nouns favours the $\text{NAC}_{\text{adj}}$ ($\beta_{\text{Measurefreq}} = -0.257$, $p_{\text{PB}} = 0.003$) as expected if we assume at least a tendency for highly grammaticalised items to be more frequent. With kind nouns, higher frequency seems to favour the $\text{PGC}_{\text{adj}}$($\beta_{\text{Kindfreq}} = 0.161$, $p_{\text{PB}} = 0.066$). However, there is no clear theoretical interpretation (see Section 2.2), and the estimate is imprecise ('not significant at $\alpha = 0.05$', see above). The effect can therefore be ignored or treated as a nuisance variable.

In Section 2.2, it was also hypothesised that there may be effects specific to classes of measure lemmas, and that they might be related to degrees of grammaticalisation of the measure noun. Transcending the effects of individual lemmas (captured in the random intercepts), the *Measureclass* second-level predictor was successfully estimated ($p_{\text{PB}} = 0.001$). Looking at the effect plot in Figure 3, it is evident that abstract non-referential physical measure nouns (such as *Gramm* 'gram' or *Liter* 'litre') with a high degree of grammaticalisation favour the $\text{NAC}_{\text{adj}}$. At the other end of the scale, nouns denoting natural portions like *Haufen* 'heap', *Bündel* 'bundle', *Schluck* 'gulp' favour the $\text{PGC}_{\text{adj}}$. These are referential nouns, confirming the hypothesis that it is prototypical of the PGC to contain two referential nouns, while the NAC only contains one (the kind noun).

I now turn to the predicted effect of cardinals as modifiers of the measure noun. Figure 4 shows that cardinals indeed influence the choice of the variant

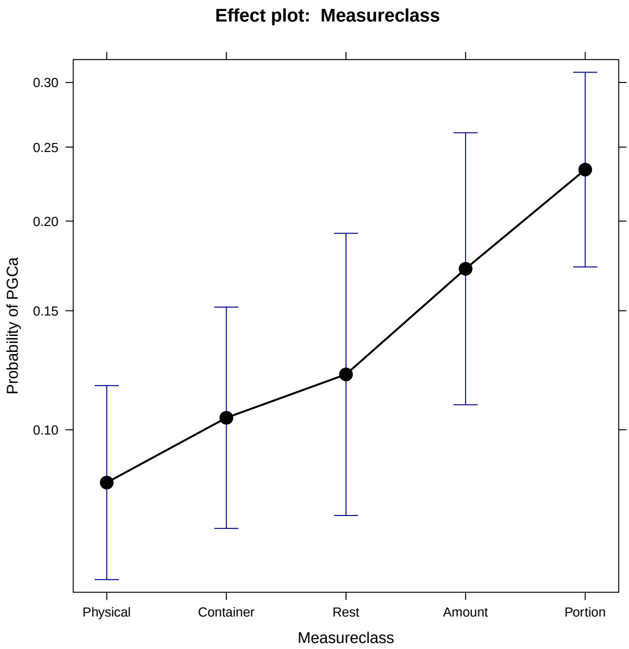


**Fig. 3:** Effect plot for the regressor *Measureclass*

($p_{\mathrm{PB}} = 0.001$), and that cardinals have a strong tendency to co-occur with the $\mathrm{NAC}_{\mathrm{adj}}$. This effect was predicted in Section 2.2.

The register-related proxy variables point into the expected direction. Increased *Badness* of the document favours the $\mathrm{NAC}_{\mathrm{adj}}$ ($\beta_{\mathrm{Badness}} = -0.165$, $p_{\mathrm{PB}} = 0.001$), and so does a lesser density of genitives ($\beta = -0.630$, $p_{\mathrm{PB}} = 0.001$). While these are poor proxies to register (and partially circular in the case of *Genitives*), this result can at least encourage future work into register effects.

The influence of *Measurecase* ($p_{\mathrm{PB}} = 0.001$) is as predicted in previous analyses (see Section 2.2). A measure noun in the dative favours the $\mathrm{PGC}_{\mathrm{adj}}$ with $\beta_{\mathrm{MeasurecaseDat}} = 0.707$ (compared to the nominative, which is on the intercept). Although *Measurecase* is a nuisance variable in the context of this study, convergence with previous work strengthens its validity.

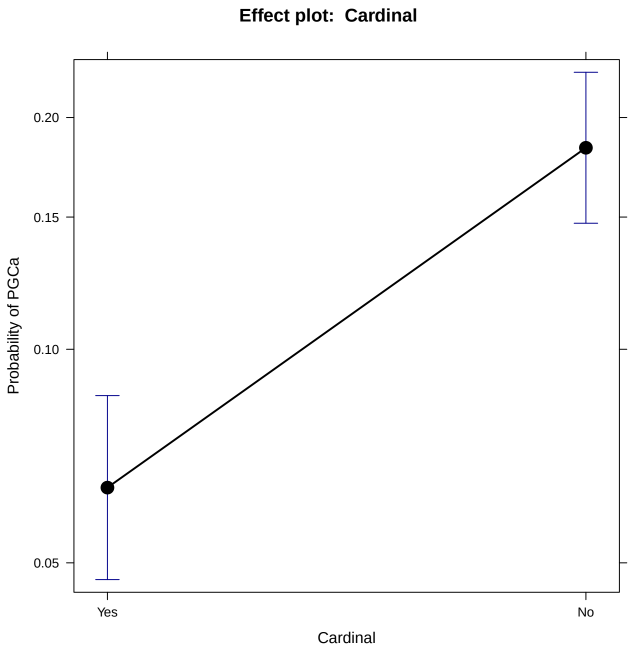


**Fig. 4:** Effect plot for the regressor *Cardinal*

# 4 Experimental validation

## 4.1 Experiment 1: forced choice

As was pointed out in Section 1, there is a strong interest in validating corpus-based findings through experiments in order to substantiate *cognitively oriented usage-based corpus linguistics* as a research programme. Therefore, this section and the next present the results of two experiments wherein I cross-checked the corpus-based findings. Both experiments use sentences containing attested MNPs from the corpus sample (embedded into simplified sentences) as stimuli. Also, the probabilities that the corpus-based model assigns to the two variants in these sentences is used as the main regressor in both studies. First, a forced-choice experiment was conducted. Participants had to chose between two sentences differing only in that one contained the $\text{NAC}_{\text{adj}}$ and the other contained the $\text{PGC}_{\text{adj}}$.

There were 24 participants (native speakers of German without reading or writing disabilities) aged 19 to 30 who were recruited from introductory lin-

| | Masculine/Neuter | Feminine |
|---|---|---|
| **high prob. for $\text{PGC}_{\text{adj}}$** | 4 sentences[a] | 4 sentences[a] |
| **low prob. for $\text{PGC}_{\text{adj}}$** | 4 sentences[a] | 4 sentences[a] |

**Table 4:** The four groups of sentences chosen as stimuli ([a]Among the 4 sentences, combinations of important factor values were made unique whenever possible.)

guistics courses at [name of university anonymised]. Although the experiment was conducted in the last four weeks of their first semester, participants had no deeper explicit knowledge of linguistics, grammar, or experimental methods. None of them had ever participated in a forced-choice experiment before. Participation was voluntary, but participants received credit in partial fulfillment of course requirements.

As stimuli, attested MNPs from the corpus study were used, but the sentences were radically simplified to avoid influences from contextual nuisance variables as much as possible. The approach is also justified because according to the theoretical assessment in Section 2.2, the choice of variants depends mostly on a very local constructional context. I sampled 16 MNPs from the corpus, and it was made sure that the simplifications and normalisations did not affect any of the regressors used in the corpus study. In the simplified sentences, the case, number, etc. of the MNP remained the same as in the attested sentence, as well as the choice of lexical material within the MNP. Eight sentences contained masculine or neuter kind nouns, the other eight contained feminine kind nouns. Furthermore, in each of the masculine/neuter and feminine groups, four sentences originally containing the $\text{NAC}_{\text{adj}}$ and four sentences originally containing the $\text{PGC}_{\text{adj}}$ were chosen. More precisely, the sentences were sampled as *highly prototypical examples* of $\text{PGC}_{\text{adj}}$ (high probability assigned by GLMM) and $\text{NAC}_{\text{adj}}$ (low probability assigned by GLMM), respectively.[19] High and low probability were defined as the top and bottom 20% of all probabilities assigned by the GLMM. Lemmas and feature combinations were made unique within each group whenever possible. The design is summarised in Table 4.

The pairs of stimuli were the sentence containing the preferred construction (according to the corpus GLMM) and a modified version containing the dispreferred construction. They were presented next to each other, and a 20 second time limit for each choice was set.[20] The position on the screen (left/right) and

**19** Remember from Section 3 that the model predicts the probability that the $\text{PGC}_{\text{adj}}$ is chosen over the $\text{NAC}_{\text{adj}}$.

**20** No participant ever exceeded the time limit.

the order of sentences were randomised for each participant. As fillers, 23 pairs of sentences exemplifying similar alternation phenomena from German morphosyntax were used. Thus, participants saw 39 pairs of sentences and 78 sentences in total. They were instructed to select from each pair of sentences the one that seemed more natural to them in the sense that they would use it rather than the other one. The experiment was conducted using *PsychoPy* (Peirce, 2007).

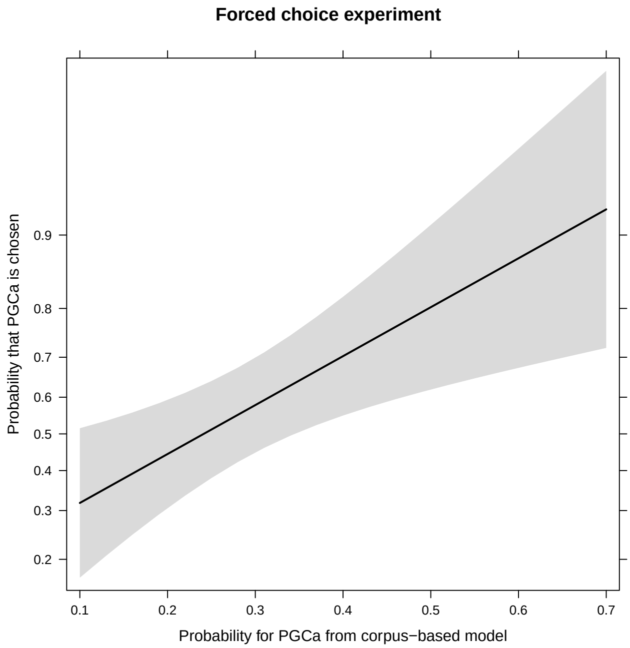


**Fig. 5:** Effect plot for the multilevel logistic regression in the forced-choice experiment: predictability of participants' choices using the probabilities derived from the corpus-based GLMM

Then, a multilevel logistic regression was specified with the probability of the $\mathrm{PGC_{adj}}$ predicted for each sentence by the corpus-based GLMM as the only fixed effect *Modelprediction*.[21] A random intercept and slope were added for the individual sentence (item) in order to catch idiosyncrasies of single sentences.

**21** The document-level variables *Badness* and *Genitives* were set to 0, which is the mean for z-transformed variables.

Also, a random intercept and slope for participants was added.[22] Coefficients were estimated with Maximum Likelihood Estimation (*lmer* function from *lme4*). The number of observations was $n = 384$.

A certain amount of the variance can be accounted for by idiosyncrasies of single sentences ($\sigma_{\text{Sentence}} = 1.785$, $\sigma^{\text{Modelprediction}}_{\text{Sentence}} = 5.996$, 16 levels).[23] Also, among participants, there are clearly different preferences ($\sigma_{\text{Participant}} = 0.781$, $\sigma^{\text{Modelprediction}}_{\text{Participant}} = 0.484$, 24 levels). On the extreme sides, one participant chose the $\text{PGC}_{\text{adj}}$ in 13 of 16 cases, and two participants only chose it in 5 of 16 cases. The regressor *Modelprediction* achieves $p_{\text{PB}} = 0.007$ ($1,000$ replications) and is estimated at 5.408 relative to an intercept of $-1.304$. The confidence interval from a parametric bootstrap ($1,000$ replications, percentile method) for the regressor is acceptable but slightly large with a lower bound of 1.626 and an upper bound of 8.397. Nakagawa & Schielzeth's pseudo-coefficients of determination are $R^2_m = 0.227$ and $R^2_c = 0.561$, which means that over 22% of the variance in the data can be explained by considering only the predictions from the corpus-based GLMM. The effect display for the single fixed regressor *Modelprediction* is given in Figure 5. The result is very clear. The higher the probability of the $\text{PGC}_{\text{adj}}$ predicted from usage-data, the more often participants chose the $\text{PGC}_{\text{adj}}$ variant in the forced-choice task. In summary, the forced choice experiment clearly succeeded in validating the results from the corpus study in as much as the preferences extracted from usage data correspond to native speakers' choices.

## 4.2 Experiment 2: self-paced reading

Finally, I report the results of a self-paced reading experiment also designed to test whether the statistical model derived from usage data (see Section 3.3) can be used to predict speakers' behaviour. It is expected that reading the less prototypical variant (the one assigned a low probability by the corpus-derived model) in a given context and with given lexical material incurs a processing overhead for the reader (Kaiser, 2013; see also Section 1). In a very similar fash-

**22** The random slopes were added to comply with Barr et al. (2013, 257) who predict *catastrophically high Type I error rates* for experimental designs with within-subject manipulations if random effects structures are not kept maximal. Notice that the model reported here was estimated with conceptually identical results with regard to the predictor of interest (*Modelprediction*) if only random intercepts were used.

**23** I use $\sigma^f_r$ to denote the standard deviation of the of the random intercepts for the fixed effect $f$ varying by random effect $r$.

ion, Divjak et al. (2016) apply the self-paced reading paradigm in the validation of corpus-based models.

I used exactly the same stimuli as in the forced choice experiments. Each participant read both the 16 sentences with the variant predicted by the corpus model and the 16 modified sentences with the variant that the corpus model does not predict.[24] To minimise repetition effects, the stimuli for each participant were separated into two blocks of 16 targets and 33 fillers per block. In the experiment, participants first read all sentences from the first block, then all sentences from the second block. It was made sure that from each target sentence pair, one sentence was assigned to the first block, and the other sentence to the other block. The assignment of members of the individual sentence pairs to the blocks was randomised for each participant individually, and so was the order within each block. The sentences from each pair of variants were kept as far apart as possible. The fillers also came in pairs such that the second block exclusively contained sentences to which participants had been exposed in the first block in slightly modified form. In total, each participant read 98 sentences. After each sentence, participants had to answer simple (non-metalinguistic) yes-no questions about the previous sentence as distractors. The distractor questions were different ones in the first and the second block. There were 38 (different) participants recruited in exactly the same manner as for the experiment reported in Section 4.1. The experiment was conducted using *PsychoPy*.

The reading times were residualised per speaker based on the reading times of all words (not just the targets) by that speaker. The adjective and the kind noun (i. e., the constituents bearing the critical case markers) were used as the target region, such as the bracketed words in the example *zwei Gläser* [*sprudelndes Wasser*] 'two glasses of sparkling water'. Outliers farther than 2 interquartile ranges from the mean logarithmised residual time were removed (64 data points), resulting in a total number of $n = 1,152$ observations. An LMM was specified with the logarithmised residual reading times as the response variable, and the probabilities derived from the corpus GLMM (*Modelprediction*) as the main regressor. Since the corpus GLMM predicts the probability of the $\mathrm{PGC}_{\mathrm{adj}}$(vs. the $\mathrm{NAC}_{\mathrm{adj}}$), it is expected that reading times are positively correlated (longer reading times) with the probability if the stimulus actually contains the $\mathrm{NAC}_{\mathrm{adj}}$, and negatively correlated (shorter reading times) if the stimulus actually contains the $\mathrm{PGC}_{\mathrm{adj}}$. To account for this, an interaction between *Mod-*

**24** Notice that lemmas and their frequencies as well as lemma classes are included as regressors in the corpus-based GLMM, and there was consequently no additional controlling of lemma frequencies, etc.

| Regressor | Coefficient | CI low | CI high | 0 in CI |
|---|---|---|---|---|
| ConstructionPGCa | 0.054 | 0.012 | 0.095 | * |
| Modelprediction | -0.006 | -0.113 | 0.110 | |
| Position | -0.005 | -0.005 | 0.004 | |
| ConstructionPGCa:Modelprediction | -0.125 | -0.234 | -0.023 | * |

**Table 5:** Fixed effect coefficient table for the LMM used to analyse the self-paced reading experiment; the intercept is 0.829

*elprediction* and *Construction* (levels *PGCa* and *NACa*) was added to the model. Furthermore, the position (1–98) of the sentence in the individual experiment (*Position* was included as a fixed effect to control for increasing reading speed during experiment runs. Random intercepts were specified for *Participant* and *Item* (the 16 sentence pairs are one *Item* each).[25]

Table 5 shows the coefficient estimates with a 95% parametric bootstrap confidence interval (1, 000 replications, percentile method). The standard deviation of the participant intercepts is $\sigma_{\text{Participant}} = 0.079$ and of the item intercepts $\sigma_{\text{Item}} = 0.037$. Comparing the full model to a model without the main regressor *Modelprediction* (and consequently also without the interaction with *Construction*) in a PB test gives $p_{\text{PB}} = 0.036$. Nakagawa and Schielzeth's pseudo-determination coefficients are $R^2_m = 0.237$ and $R^2_c = 0.346$.

The overall model quality is acceptable, and the effect plot for the main effect is shown in Figure 6. The estimate for the sentences with $\text{NAC}_{\text{adj}}$ is obviously imprecise, although pointing into the right direction (longer reading times when the corpus model predicts the $\text{PGC}_{\text{adj}}$). There is a clearer effect in the sentences with $\text{PGC}_{\text{adj}}$, which is also confirmed by the 'significant' results from the bootstrapped confidence intervals (see Table 5) and from the PB test reported above. The $\text{PGC}_{\text{adj}}$ brings about an increased reading time ($\beta_{\text{ConstructionPGCa}} = 0.054$), which is plausible because it is the much rarer construction (see Section 3). However, if it occurs in a prototypical context and with prototypical lexical material, reading times drop ($\beta_{\text{ConstructionPGCa:Modelprediction}} = -0.125$). This can be seen in the downward slope of curve in the right panel of Figure 6. This fits into the general picture in as much as the construction with the lower frequency might be

**25** I tried random slopes in order to keep the random effect structure maximal (Barr et al., 2013), but it was impossible to get the algorithm to converge due to the added complexity of the interaction.

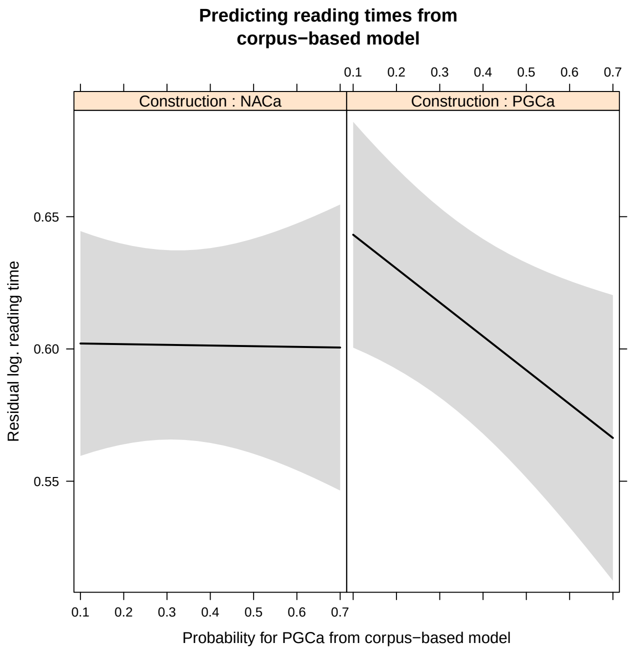


**Fig. 6:** Effect plot for the LMM in the self-paced reading experiment: modeling participants' residualised log reading times on the probabilities given by the corpus-based GLMM

developing towards a more sharply defined prototype.[26] Conversely, the $\mathrm{NAC_{adj}}$ (like the NAC in general) might be the highly frequent default which does not incur reading time penalty, even if it is not the optimal choice in the given context and with the given lexical material. This interpretation will be considered, among other things, in the next and final section.

# 5 Conclusions

I have shown! Oh, yes, I have shown...

**26** In this context, it should be remembered from Section 3.1 that even the $\mathrm{PGC_{det}}$ is much rarer that the $\mathrm{NAC_{bare}}$ ($17,252$ vs. $315,635$ occurrences in the auxiliary corpus samples).

# References


Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics* 11. 295–328.

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova & Tore Nesset. 2013. Making choices in russian: pros and cons of statistical methods for rival forms. *Russian Linguistics* 37. 253–291.

Baayen, R. Harald, Cyrus Shaoul, Jon Willits & Michael Ramscar. 2016. Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience* 31(1). 106–128.

Barker, Chris. 1998. Partitives, double genitives and anti-uniqueness. *Natural Language and Linguistic Theory* 16(4). 679–717.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3). 255–278.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Bhatt, Christa. 1990. *Die syntaktische struktur der nominalphrase im deutschen*. Tübingen: Narr.

Biemann, Chris, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski & Torsten Zesch. 2013. Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics* 28(2). 23–60.

Bildhauer, Felix & Roland Schäfer. 2016. Automatic classification by topic domain for meta data generation, web corpus evaluation, and corpus comparison. In Paul Cook, Stefan Evert, Roland Schäfer & Egon Stemle (eds.), *Proceedings of the 10th web as corpus workshop (WAC-X)*, 1–6. Association for Computational Linguistics.

Brems, Lieselotte. 2003. Measure noun construction: An instance of semantically-driven grammaticlization. *International Joünal of Corpus Linguistics* 8(2). 283–312.

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

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213.

Bresnan, Joan & Jennifer Hay. 2008. Gradient grammar: An effect of animacy on the syntax of 'give' in New Zealand and American English. *Lingua* 118. 245–259.

Carpenter, Bob, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li & Allen Riddell. 2017. Stan: A probabilistic programming language. *Journal of Statistical Software* 76(1). 1–32.

De Clerck, Bernard & Lieselotte Brems. 2016. Size nouns matter: A closer look at mass(es) of and extended uses of SNs. *Language Sciences* 53. 160–176.

Divjak, Dagmar. 2016. Four challenges for usage-based linguistics. In Jocelyne Daems, Eline Zenner, Kris Heylen, Dirk Speelman & Hubert Cuyckens (eds.), *Change of paradigms – new paradoxes – recontextualizing language and linguistics*, 297–309. Berlin/Boston: De Gruyter Mouton.

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

Divjak, Dagmar, Antti Arppe & R. Harald Baayen. 2016. Does language-as-used fit a self-paced reading paradigm? In Tanja Anstatt, Anja Gattnar & Christina Clasmeier (eds.), *Slavic languages in psycholinguistics*, 52–82. Tübingen: Narr Francke Attempto.

Eschenbach, Carola. 1994. Maßangaben im Kontext - Variationen der quantitativen Spezifikation. In Sascha W. Felix, Christopher Habel & gert Rickeit (eds.), *Kognitive Linguistik – Repräsentationen und Prozesse*, 207–228. Opladen: Westdeutscher Verlag.

Fleischer, Jürg & Oliver Schallert. 2011. *Historische Syntax des Deutschen : eine Einführung*. Tübingen: Narr.

Fox, John. 2003. Effect displays in R for Generalised Linear Models. *Journal of Statistical Software* 8(15). 1–27.

Fox, John & Georges Monette. 1992. Generalized collinearity diagnostics. *Journal of the American Statistics Association* 87. 178–183.

Freedman, David. 1999. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics* 27(4). 1119–1140.

Gabry, Jonah & Ben Goodrich. 2016. *rstanarm: Bayesian applied regression modeling via stan*. R package version 2.12.1.

Gallmann, Peter & Thomas Lindauer. 1994. Funktionale Kategorien in Nominalphrasen. *Beiträge zur Geschichte der deutschen Sprache* 116.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari & Donald B. Rubin. 2014. *Bayesian data analysis*. Boca Raton: Chapman & Hall 3rd edn.

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, Andrew & Cosma Rohilla Shalizi. 2013. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* 66. 8–38.

Goethem, Kristel Van & Philippe Hiligsmann. 2014. When two paths converge: Debonding and clipping of Dutch 'reuze'. *Journal of Germanic Linguistics* 26(1). 31–64.

Goethem, Kristel Van & Matthias Hüning. 2015. From noun to evaluative adjective: Conversion or debonding? Dutch top and its equivalents in German. *Journal of Germanic Linguistics* 27(4). 365–408.

Grestenberger, Laura. 2015. Number marking in German measure phrases and the structure of pseudo-partitives. *Journal of Comparative Germanic Linguistics* 18. 93–138.

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

Gries, Stefan Th. 2015b. The role of quantitative methods in cognitive linguistics: corpus and experimental data on (relative) frequency and contingency of words and constructions. In Jocelyne Daems, Eline Zenner, Kris Heylen, Dirk Speelman & Hubert Cuyckens (eds.), *Change of paradigms - new paradoxes: recontextualizing language and linguistics*, 311–325. Berlin/New York: De Gruyter Mouton.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Co-varying collexemes in the into-causative. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture, and mind*, 225–236. Stanford, CA: CSLI.

Halekoh, Ulrich & Søren Højsgaard. 2014. A Kenward-Roger approximation and parametric bootstrap methods for tests in linear mixed models – the R package pbkrtest. *Journal of Statistical Software* 59(9). 1–30.

Hankamer, Jorge & Line Mikkelsen. 2008. Definiteness marking and the structure of Danish pseudopartitives. *Journal of Linguistics* 44(2). 317–346.

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

Kaiser, Elsi. 2013. Experimental paradigms in psycholinguistics. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 135–168. Cambridge University Press.

Kilgarriff, Adam. 2006. Googleology is bad science. *Computational Linguistics* 33(1). 147–151.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. 2014. The Sketch Engine: ten years on. *Lexicography* 1–30.

Klein, Wolf-Peter. 2009. Auf der Kippe? Zweifelsfälle als Herausforderung(en) für Sprachwissenschaft und Sprachnormierung. In Marek Konopka & Bruno Strecker (eds.), *Deutsche Grammatik – Regeln, Normen, Sprachgebrauch*, Berlin: De Gruyter.

Koptjevskaja-Tamm, Maria. 2001. "A piece of the cake" and "a cup of tea": partitive and pseudo-partitive nominal constructions in the Circum-Baltic languages. In Östen Dahl & Maria Koptjevskaja-Tamm (eds.), *The Circum-Baltic languages: Typology and contact*, vol. 2, 523–568. Amsterdam and Philadelphia: John Benjamins.

Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt. 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 & Daniel Tapias (eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC '10)*, 1848–1854. Valletta, Malta: European Language Resources Association (ELRA).

Labov, William. 1966. *The social stratification of english in new york city*. Washington, DC: Center for Applied Linguistics.

Labov, William. 1969. Contraction, deletion, and inherent variability of the english copula. *Language* 45(4). 715–762.

Lehmann, Erich L. 1993. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistics Association* 88. 1242–1249.

Lehmann, Erich L. 2011. *Fisher, neyman, and the creation of classical statistics*. New York, NY: Springer.

Levshina, Natalia. 2016. When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2). 235–268.

Löbel, Elisabeth. 1986. Apposition in der Quantifizierung. In Armin Burkhardt & Karl-Hermann Körner (eds.), *Pragmantax. Akten des 20. Linguistischen Kolloquiums Braunschweig 1985*, .

Löbel, Elisabeth. 1989. Q as a functional category. In Christa Bhatt, Elisabeth Löbel & Claudia Schmidt (eds.), *Syntactic phrase structure phenomena*, 133–158. Amsterdam, Philadelphia: Benjamins.

Maxwell, Scott E. & Harold D. Delaney. 2004. *Designing experiments and analyzing data: A model comparison perspective*. Mahwa, New Jersey, London: Taylor & Francis.

Mayo, Deborah G. 2013. The error-statistical philosophy and the practice of Bayesian statistics: Comments on Gelman and Shalizi: 'Philosophy and the practice of Bayesian statistics'. *British Journal of Mathematical and Statistical Psychology* 66. 57–64.

Milin, Petar, Dagmar Divjak, Strahinja Dimitrijević & R. Harald Baayen. 2016. Towards cognitively plausible data science in language research. *Cognitive Linguistics* 27(4). 507–526.

Müller, Sonja. 2014. Zur Anordnung der Modalpartikeln "ja" und "doch": (In)stabile Kontexte und (non)kanonische Assertionen. *Linguistische Berichte* 238. 165–208.

Nakagawa, Shinichi & Holger Schielzeth. 2013. A general and simple method for obtaining $R^2$ from generalized linear mixed-effects models. *Methods in Ecology and Evolution* 4(2). 133–142.

Nesset, Tore & Laura A. Janda. 2010. Paradigm structure: Evidence from russian suffix shift. *Cognitive Linguistics* 21(4). 699–725.

Peirce, Jonathan W. 2007. Psychopy – Psychophysics software in Python. *Journal of Neuroscience Methods* 162(1–2). 8–13.

Perezgonzalez, Jose D. 2015. Fisher, Neyman-Pearson or NHST? A tutorial for teaching-data testing. *Frontiers in Psychology* 6(223). 1–11.

R Core Team. 2014. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. http://www.R-project.org/.

Ramscar, Michael & Robert F. Port. 2016. How spoken languages work in the absence of an inventory of discrete units. *Language Sciences* 53. 58–74.

Rutkowski, Paweł. 2007. The syntactic structure of grammaticalized partitives (pseudo-partitives). In Tatjana Scheffler, Joshua Tauberer, Aviad Eilam, & Laia Mayol (eds.), *Proceedings of the 30th annual penn linguistics colloquium*, vol. 1 (University of Pennsylvania Working Papers in Linguistics 13), 337–350. Philadelphia: Pennsylvania Graduate Linguistics Society.

Schachtl, Stefanie. 1989. Morphological case and abstract case: Evidence from the German genitive construction. In Christa Bhatt, Elisabeth Löbel & Claudia Schmidt (eds.), *Syntactic phrase structure phenomena*, 99–112. Amsterdam, Philadelphia: Benjamins.

Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen & Andreas Witt (eds.), *Proceedings of Challenges in the Management of Large Corpora 3 (CMLC-3)*, UCREL Lancaster: IDS.

Schäfer, Roland. 2016 aop. Prototype-driven alternations: The case of german weak nouns. *Corpus Linguistics and Linguistic Theory* ahead of print.

Schäfer, Roland, Adrien Barbaresi & Felix Bildhauer. 2013. The good, the bad, and the hazy: Design decisions in web corpus construction. In Stefan Evert, Egon Stemle & Paul Rayson (eds.), *Proceedings of the 8th web as corpus workshop (wac-8)*, 7–15. Lancaster: SIGWAC.

Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*, 486–493. Istanbul, Turkey: European Language Resources Association (ELRA).

Schäfer, Roland & Felix Bildhauer. 2013. *Web corpus construction* (Synthesis Lectures on Human Language Technologies). San Francisco: Morgan and Claypool.

Schäfer, Roland & Ulrike Sayatz. 2014. Die Kurzformen des Indefinitartikels im Deutschen. *Zeitschrift für Sprachwissenschaft* 33(3). 215–250.

Schäfer, Roland & Ulrike Sayatz. 2016. Punctuation and syntactic structure in "obwohl" and "weil" clauses in nonstandard written German. *Written Language and Literacy* 19(2). 215–248.

Selkirk, Elisabeth O. 1977. Some remarks on noun phrase structure. In Peter W. Culicover, Thomas Wasow & Adrian Akmajian (eds.), *Formal syntax: Papers from the MSSB-UC Irvine conference on the formal syntax of natural language, Newport Beach, California*, 285–316. New York: Academic Press.

Senn, Stepen. 2011. You may believe you are a bayesian but you are probably wrong. *Rationality, Markets, and Morals* 2. 48–66.

Stickney, Helen. 2007. From pseudopartitive to partitive. In Alyona Belikova, Luisa Meroni & Umeda Mari (eds.), *Proceedings of the 2nd conference on generative approaches to language acquisition North America (GALANA)*, 406–415. Somerville.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Malden/Oxford: Wiley-Blackwell.

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen & hans van Halteren. 2013. Choosing alternatives: Using Bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9(2). 227–262.

Vos, Riet. 1999. *A grammar of partitive constructions* (Tilburg dissertation in language studies). Tilburg: Tilburg University.

Zeldes, Amir. to appear. The case for caseless prepositional constructions with "voller" in German. In Hans C. Boas & Alexander Ziem (eds.), *Constructional approaches to argument structure in German* (Trends in Linguistics: Studies and Monographs), Berlin: De Gruyter.

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.

Zuur, Alain F., Elena N. Ieno & Chris S. Elphick. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution* 1(1). 3–14.