Roland Schäfer*

Abstractions and exemplar effects in the German measure noun phrase alternation

Abstract: In this paper, a case alternation in German measure noun phrases is examined from the perspective of cognitively oriented corpus linguistics. In certain pseudo-partitive constructions, the embedded kind-denoting noun either agrees in its case with the measure head noun (*eine Tasse guter Kaffee* 'a cup of good coffee') or it stands in the genitive (*eine Tasse guten Kaffees*). My analysis is formulated in terms of Prototype Theory. I assume that the choice of alternants is influenced by non-alternating neighbouring constructions representing the prototypes more directly. I argue that the frequencies with which individual lemmas occur in these prototypical non-alternating constructions partially predict which alternant is chosen and that the genitive alternant is prototypical of strongly grammaticalised measure nouns. I present a large-scale corpus study using the DECOW corpus to support this theory. In the statistical analysis, I compare Maximum Likelihood estimators to Bayesian estimators (recently proposed for similar studies), showing that results *predictably* do not differ. Finally, two experiments (forced choice and self-paced reading) are reported. The usage-based findings show a clear correlation with the behaviour of cognitive agents in both experiments. The present study therefore contributes to the well-established field of research into alternations using corpus and experimental methods.

Keywords: corpus methods and experimental methods, alternations, hierarchical models, pseudo-partitives, German

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

1 Prototypes, and exemplars, and corpora

This paper deals with a morpho-syntactic alternation between two constructions that occurs only in a very specific type of measure noun phrase 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. The study of alternations has a long history in cognitively oriented corpus linguistics (for example, Bresnan et al., 2007; Bresnan & Hay, 2008; Bresnan & Ford, 2010; Divjak & Arppe, 2013; Gries, 2015a; Nesset & Janda, 2010). This area of research is based on the assumption that language is a probabilistic phenomenon (Bresnan, 2007) where alternants are chosen neither deterministically nor fully at random. Instead, multifactorial models are constructed which incorporate influencing factors from diverse levels, including contextual factors. The estimation of the model coefficients quantifies the influence that the factors have on the probability that either alternant is chosen. There are two fundamental issues to consider. First, there is a question of how corpus data relates to experimental findings, which provide more direct evidence of mental representations and processes. Second, the appropriate modelling of such results in cognitive linguistics is a key issue.

As for the first issue, there has always been an interest in correlating probabilistic generalisations extracted from corpus data with results from experimental work (for example, Arppe & Järvikivi, 2007; Bresnan et al., 2007; Bresnan & Ford, 2010; Divjak & Gries, 2008; Divjak et al., 2016; Ford & Bresnan, 2013). This is often called a *validation* of the corpus-derived findings, but Divjak (2016, 303) rightly criticises this choice of words "because it creates the impression that behavioral experimental data is inherently more valuable than textual data," citing Tummers et al. (2005), who state that a corpus is "a sample of spontaneous language use that is (generally) realized by native speakers." However, as Dąbrowska (2016, 486–487) convincingly argues, this does not imply that we can in some way "deduce mental representations from patterns of use," i. e., from corpus data. It would be highly surprising if this were possible, and the same holds for experimental methods. Nobody assumes that we can inductively infer mental representations from experiments, which – as opposed to corpus studies – even allow for direct access to the cognitive agent and offer much better possibilities to control experimental conditions and nuisance variables. Rather, under the standard approach, a theory of cognitive representation is pre-specified. Then, predictions are derived from this theory *before* the experiment or the corpus study is conducted to *test* the theory. While the same approach is by and large applicable to corpus data.

Discussion of non-convergence here.

Turning to the second issue, This entails that a variant of prototype theory with features is assumed (Rosch, 1978). Prototype theory is well suited for modeling constructional choices, but it is just one of several similarity-based theories of classification, the most prominent other framework being exemplar theory (Medin & Schaffer, 1978; Hintzman, 1986; see Storms et al., 2000 for a comparison of the theories in different experimental settings). Prototype theory and exemplar theory model essentially the same types of effects but differ significantly in whether they assume higher-level abstractions in the form of single maximally prototypical exemplars or their features (prototype theory) or assume that categories emerge through the storage of many exemplars and similarity classification on those exemplars (exemplar theory). Parallel to Langacker here

As Barsalou (1990) already showed, however, prototype and exemplar theory model the same types of effects and are informationally equivalent. Consequently, experiments which favour one theory over the other use procedural behaviour of subjects, for example the speed of category retrieval, and not mere output data. In very early experiments, Posner & Keele (1968) showed, for example, that highly prototypical unseen exemplars were categorised more easily by subjects compared to less prototypical ones, even if these were included in the learning data. Since corpus data only show artefacts of production events and we have no experimental access to the speaker's or writer's performance, one should be sceptical whether corpus analysis alone could ever decide which theory of mental representation is more suitable (see also Gries, 2003, 22 and the Dąbrowska, 2016, 486–487 quote above).

In cognitive science, it is mostly accepted that exemplar theories have greater explanatory power (Vanpaemel, 2016, 184), and that abstraction is only needed marginally, if at all.[1] Still, various attempts have been made over the past decades to settle the dispute between abstraction-based models (models with rules or prototypes) and exemplar models or to find models which unite the

1 The hard empirical evidence in favour of exemplar models is substantial. For example, in Hahn et al. (2010), the authors show that subjects even use exemplar similarity over abstract knowledge even when they are given very simple explicit rules. This is highly relevant because most other studies focus on the learning of implicit rule-based knowledge, which involves many auxiliary assumptions in actual experiments (Hahn et al., 2010, 2). On the other hand, there is evidence that neither theory is fully adequate to model humans' capabilities to form categories. For example, Conaway & Kurtz (2016) show that reference point approaches to category learning such as prototype and exemplar theory fail to explain certain experimental results where subjects learn to generalise beyond the input in a way that cannot be explained by similarity.

two extremes. Vanpaemel & Storms (2008); Lee & Vanpaemel (2008) proposed the *varying abstraction model* (VAM) which "attempt[s] to balance economy and informativeness" (Lee & Vanpaemel, 2008, 745), treating models with full abstraction (prototypes) and no abstraction at all (radical exemplar theory) as special cases of a model which allows for both abstraction and exemplar effects. The mixture model of categorisation (MMC) by Rosseel (2002) is a model with abstraction in the form of hierarchical clusters of exemplars, and these clusters of objects are characterised by a probability distribution over their features, and categorising new objects is a process of estimating the probability of this object being from one of the clusters. Griffiths et al. (2009) go further and present a computational model based on the hierarchical Dirirchlet process which is able to chose the appropriate complexity of representation for a given category. However, despite these (and more) attempts to , (Vanpaemel, 2016, 183–184) describes the state of affairs between adherents of neo-prototype theory (such as Minda & Smith, 2001, 2002) and exemplar theory as a stalemate, and suggests yet another approach to solve the divide.

In cognitive linguistics, Divjak & Arppe (2013) is a very rare example of a paper where such issues are taken up. Their corpus-based approach shows "one way of systematically analyzing usage data as contained in corpora to yield a scheme, compatible with usage-based theories of language, by which the assumptions of both the prototype and exemplar theories can be operationalized" (Divjak & Arppe, 2013, 267). Their approach to implementing a varying abstraction model (Divjak & Arppe, 2013, 254–260) is based on hierarchical clustering of annotated properties of sentences. They hierarchically cluster sentences containing Russian verbs of trying. Then, they single out one sentence from each cluster which scores the highest probability for any of the six *try* verbs according to a polytomous regression model estimated on the same data. The clusters are interpreted as intermediate-level exemplar-derived abstractions of typical contexts for these high-probability verbs (typically more than one cluster for each verb; Divjak & Arppe, 2013, 255–256). This is pioneering work, but the crucial difference between data-driven corpus-based analyses and experiments in cognitive science (they use Verbeemen et al., 2007 as their reference) is that cognitive experimental research is based on experiments where subjects produce category assignments, and in corpus studies, the categories and category membership is determined purely from the data.

One thing that should be kept in mind is that most of the research on categorisation in cognitive science uses experiments with highly simplified stimuli and very simple tasks. It is remarkable in this context that Voorspoels et al. (2011) consider the experimental task reported by them – assigning typicality scores to nouns from the domains of *animals* and *artefacts* to categories like *bird*,

fish, *clothing*, or *tools* – a study of "superordinate natural language categories, whereas most evidence supporting exemplar representations has been found in artificial categories of a more subordinate level" (Voorspoels et al., 2011, 1013). Linguists (not just of the cognitive variety) are usually interested in much more complex high-level categories and use large and complex feature sets, especially in (morpho-)syntax.[2]

As a consequence, formal models of categorisation in cognitive linguistics have not been spelled out at the same level of mathematical detail as in cognitive science. However, the fact that linguists deal with real data should be seen as an advantage in that it greatly improves the *external validity* of studies, i. e.,the generalisability of the findings. Typical artificial tasks in cognitive science have been criticised exactly for the lack of external validity (e. g.,Murphy, 2003).

I propose that the primary focus should be on determining which factors influence choices made by speakers. As (Gries, 2003, 22) put it with reference to classic prototype theory, "even if the form of analysis does not translate into statements on mental representations, the high predictive power [...] shows that the cognitive factors underlying the choice of construction have been identified properly and weighted in accordance with their importance for actual usage." A major advancement since this early time has been the focus on which effects are at least more plausibly modelled as prototype/abstraction and exemplar effects, and how observed data fits the two approaches (see, for example, Divjak & Arppe, 2013). This paper contributes this discussion.

2 Notice that recently, approaches have emerged which solve at least some problems by abandoning linguistic high-level features altogether (Baayen et al., 2016; Ramscar & Port, 2016).

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 to 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-denoting noun. I call the kind-denoting noun the *kind noun* and the quantity-denoting noun the *measure noun*. For illustration purposes, in the English *a glass of good wine*, *glass* is the measure noun and *wine* is the kind noun. Measure nouns can be all sorts of nouns which denote a quantity (such as *litre* or *amount*) but also those denoting containers, collections, etc. (such as *glass* or *bucket*). Like Brems (2003, 284), I consider nouns as measure nouns "which, strictly speaking, do not designate a 'measure', but display a more nebulous potential for quantification" (also Koptjevskaja-Tamm, 2001, 530, and Rutkowski, 2007, 338).

2.1.1 Core structures related to the alternation

Three different syntactic configurations within MNPs need to be distinguished, and the case alternation occurs only in one of them. It occurs only when the kind noun is modified by an attributive adjective and there is no determiner, as in (1). Superficially, the sentences are functionally and semantically equivalent either with the kind noun in the genitive (1a) or in the same case as the measure noun, an accusative in the case of (1b).[3]

(1) a. Wir trinken [[ein Glas]$_{\text{Acc}}$ [guten Weins]$_{\text{Gen}}$]$_{\text{Acc}}$.
we drink a glass good wine
We drink a glass of good wine.

b. Wir trinken [[ein Glas]$_{\text{Acc}}$ [guten Wein]$_{\text{Acc}}$]$_{\text{Acc}}$.

3 Some descriptive and normative grammars take stronger positions with regard to the acceptability of the two options. See Hentschel (1993) and Zimmer (2015) for analyses of the sometimes absurd stances taken in grammars of German. As will be shown (especially in Section 4.1), there might be preferences, but we cannot assume either construction to be unacceptable.

This specific configuration has to be seen in the context of two other configurations, to which I turn now. First, 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 Gerstenberger, 2015).[4] Here, the kind noun is in the genitive, and I refer to the construction in (2) as the *Pseudo-partitive Genitive Construction* (PGC).

(2) Wir trinken [[ein Glas]$_{\text{Acc}}$ [dieses Weins]$_{\text{Gen}}$]$_{\text{Acc}}$.
we drink a glass this wine
We drink a glass of this wine.

Second, if the kind noun is bare – i. e., if it comes neither with a determiner nor a modifying adjective – it is uninflected as in (3a), and the genitive as seen in the PGC is not acceptable, see (3b).

(3) a. Wir trinken [[ein Glas]$_{\text{Acc}}$ [Wein]$_{?}$]$_{\text{Acc}}$.
we drink a glass wine
We drink a glass of wine.

b. * Wir trinken [[ein Glas]$_{\text{Acc}}$ [Weins]$_{\text{Gen}}$]$_{\text{Acc}}$.

This construction is usually classified as a *Narrow Apposition Construction* (Löbel, 1986), henceforth NAC.[5] Notice that the unavailability of the genitive on the kind noun follows independently from a 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; see also Eisenberg, 2013, 160).

It is difficult to determine whether the bare kind noun in the narrow apposition construction as in (3a) bears no case at all, a generic case, or agrees in case

4 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).

5 The construction as in (3a) 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.

with the measure noun.[6] When there is an adjective as in (1b), the embedded kind NP clearly agrees in case, but due to the overall absence of markers of case in the singular, bare nouns mostly show no indication of their case. The only nouns which do have case markers in the singular are the so-called weak nouns (Köpcke, 1995; Schäfer, 2016), which have something like a non-nominative *-en* marker in the singular. Unfortunately, there are no genuine mass nouns among the weak nouns. However, a few of them can be coerced into a mass noun, such as *Hase* 'rabbit' (meaning 'rabbit meat'). It then appears as if the uninflected form is preferred, but the inflected form is not excluded. In (4a), the clearly acceptable form *Hase* can only be a nominative singular or caseless. In (4b), the form *Hasen* could be an accusative, dative, or genitive.

(4) a. Niemand will [ein Stück [Hase]$_{\text{Nom/caseless}}$]$_{\text{Acc}}$ essen.
nobody wants a piece rabbit eat
Nobody wants to eat a piece of rabbit.

b. ? Niemand will [ein Stück [Hasen]$_{\text{Acc/Dat/Gen}}$]$_{\text{Acc}}$ essen.
nobody wants a piece rabbit eat

Even a full unacceptability of (4b) would not be conclusive, however, as a possible aversion of speakers towards the case-marked form might be due to the fact that it is at least potentially also a genitive, in which case the constraint against bare genitive nouns would apply. With plural kind nouns, the obligatory marking of the dative plural with *-en* might provide some clues. However, plural kind nouns do not behave like singular ones in measure phrases anyway, as will be argued in Section 2.1.2. Also, judgements vary between (5a) and (5b), and both are found in corpora.[7]

(5) a. mit [zwei Säcken [Äpfel]$_{\text{Nom/Acc/Gen}}$]$_{\text{Dat}}$
with a sack apples
with a sack of apples

b. mit [zwei Säcken [Äpfeln]$_{\text{Dat}}$]$_{\text{Dat}}$
with a sack apples

Descriptive grammars seem to favour an analysis in terms of caselessness (for example, Zifonun et al., 1997, 1981). The hazy picture of the case of bare kind

6 I would like to thank one of the reviewers for pointing this out.

7 For example, the variant in (5a) with the lemmas *Sack* and *Apfel* occurs four times, and the one in (5b) twice in the very large DECOW corpus (see Section 3.1). Similarly, for [*Kiste* [*Äpfel*]$_{\text{Nom/Acc/Gen}}$]$_{\text{Dat}}$ 'box of apples' (no case identity), I find three examples, and [*Kiste* [*Äpfeln*]$_{\text{Dat}}$]$_{\text{Dat}}$ (clearly marked case identity), I find six examples.

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} (3a) *Glas Wein*	NAC_{adj} (1b) *Glas guten Wein*	—
pseudo-partitive genitive	—	PGC_{adj} (1a) *Glas guten Weins*	PGC_{det} (2) *Glas dieses Weins*

Table 1: NAC and PGC constructions in different NP structures with examples and references to full example sentences

NPs is most likely due to the fact that case is so rarely marked on German nouns, where case is marked mostly on determiners and to some degree adjectives. The uncertainty in the few cases where case can be marked (weak nouns in the singular and dative plurals) would thus be a direct consequence of the fact that the construction is not very specific with respect to the case of the kind noun.

To summarise, the case patterns in the NAC and in the PGC (depending on the structure of the kind NP) is given in Table 1. I call the narrow apposition construction with a bare kind noun the NAC_{bare} and the partitive genitive with a determiner in the kind NP the PGC_{det}. For the alternants with an adjective but no determiner in the kind noun phrase I use the terms NAC_{adj} and PGC_{adj}. In principle, this paper is about the middle column of Table 1, i. e., the syntactic configuration in which two different case patterns are acceptable. However, in Section 2.2, the outer columns (NAC_{bare} and PGC_{det}) will still play a major role when the factors controlling the alternation are discussed.

2.1.2 Neighbouring cases

I now turn to some more subtle issues related to the measure noun case alternation in order to delimit the scope of the study. First, there is a claim 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) and Zimmer (2015). Sentence (6) shows a putative generic nominative on the kind noun inside an accusative MNP.

(6) * Wir trinken [[eine Tasse]$_{Acc}$ [heißer Kaffee]$_{Nom}$]$_{Acc}$.

It was shown empirically in Hentschel (1993) that such sentences 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 corpus linguistic methods (Section 3), and I consequently do not discuss them further.

Second, we find that, if the kind noun is a plural count noun as in *ein Sack kleine Äpfel* (NAC_{adj}) or *ein Sack kleiner Äpfel* (PGC_{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 (794 of 861 cases, or over 92%, cf. Section 3) that the alternation cannot be analysed in the same way as in the singular. 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 in Sections 3 and 4.

Third, some measure nouns have been grammaticalised in a way that they always appear in their non-inflected form. They are typical measure nouns like *Gramm* 'gram', *Pfund* 'pound' or *Prozent* 'percent', which have no normal plural forms at all.[8] I treat these cases like other measure nouns because they enter into both the NAC_{adj} as in (7a) and the PGC_{adj} as in (7b). In Section 2.2, however, degrees of grammaticalisation as a factor influencing the alternation will be discussed prominently.

(7) a. zwei Gramm brauner Zucker

b. zwei Gramm braunen Zuckers
two gram brown sugar
two grams of brown sugar

Finally, 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 Tagliamonte, 2012). In the case at hand, the variable might be something like *measurement of quantities of substances and collections*. I argue that it is fully justified to focus narrowly on NAC_{adj} and PGC_{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. There is an analytic (pseudo-)partitive with *von* 'of'. It is only available as an alternative to the PGC if the kind noun phrase contains a (definite or indefinite) determiner as in (8).

8 Plurals like *Pfunde* and *Prozente* have special meanings and have very restricted uses, mostly in idiomatic expressions such as *Pfunde verlieren* 'lose pounds' and *Prozente machen* 'make a profit'. They cannot be used in normal MNPs.

(8) 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 NAC_{adj} and the PGC_{adj}. In fact, there is not a single context in which they can be interchanged.

Additionally, we find constructions with *voll* (*von*)/*voller* 'full (of)' and *mit* 'with' are available, see (9).

(9) 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) and the construction with *mit* 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 used only with measure nouns denoting containers, where the whole NP typically refers to the container and never to the quantity contained in it. They are 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 NAC_{adj} or PGC_{adj} in most contexts, and they are thus clearly not just more ways of saying the same thing. It is therefore reasonable to focus on the two well-defined alternants.

This concludes the descriptive overview of the phenomenon. I have demonstrated 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 more theory-oriented discussion of the alternation in the next section.

2.2 Prototype and exemplar effects

This section briefly reviews existing analyses of the $\mathrm{PGC}_{\mathrm{adj}}$ vs. $\mathrm{NAC}_{\mathrm{adj}}$ alternation and related issues. I also develop my own analysis and the appropriate hypotheses for the empirical studies presented in Sections 3 and 4.
.[9]

9 A reviewer mentioned that she or he had the impression that dialectal variation is also a factor influencing the alternation. This is definitely true as some dialects (such as Alemannic) tend to have no genitive at all. While this is an interesting aspect for further research, the main obstacle is that there are no corpora of German which are both large enough and annotated with reliable meta data about regional variants. In the annotation of the corpus study, documents obviously written in a regional variant were excluded.

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, and Schäfer, 2015, as well as Biemann et al., 2013, and 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 external 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 proofreaders who must adhere to normative guidelines might favour one alternative or the other explicitly. 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), Schäfer & Sayatz (2014), Schäfer & Sayatz (2016), and Zimmer (2015). Therefore, DECOW is a valid choice for this study.

10 The COW corpora (Dutch, English, French, German, Spanish, Swedish) are made available for free at https://www.webcorpora.org. At the time of this writing, a newer 2016 version DECOW16 has 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 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 non-replicability of results, as Kilgarriff (2006) pointed out more than ten years ago. Careless use of search engine results is still found, however, see for example De Clerck & Brems (2016, 171–175).

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

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 process in order to obtain such a sample, which consisted of the following steps:

i. generating a list of the one hundred most frequent mass nouns,
ii. deriving a list of all measure nouns with which the mass nouns co-occur in the $\mathrm{NAC}_{\mathrm{bare}}$, and
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 noun downwards, selecting the first one hundred mass nouns that occurred in the list.[13] Mass nouns were defined as concrete nouns which denote a substance in the broad sense, combine with uninflected mass quantifiers such as *viel* 'much' and *wenig* 'little' (*viel Bier* 'much beer'), and form only sortal and unit plurals (such as the plural *Biere* 'types of beer' or 'glasses of beer'). 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* 'jewellery', which is the 3,054th most frequent noun in the original frequency list.

This list of mass nouns was used in step (ii) to derive 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 $\mathrm{N_1N_2}$ where $\mathrm{N_2}$ 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.

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

In step (iii), each of these 2,365 noun–noun combinations was queried in the target constructions (PGC_{adj} and NAC_{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 (often referred to as a *Zipfian* distribution), I used all hits for nouns with a frequency up to 100 and a sample of 100 of all those with higher frequency. 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 NAC_{adj} cannot be distinguished from the PGC_{adj}. Given the careful 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]

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 NAC_{bare} and the PGC_{det} with a determiner will be modelled as factors influencing the alternation. Therefore, all noun-noun pairs from the process described above were also queried in the two non-alternating constructions, resulting in 17,252 hits for the PGC_{det} and 315,635 hits for the NAC_{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 first 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 PGC_{det} and the NAC_{bare}. They were calculated from the auxiliary samples described at the end of Section 3.1 as a log-transformed quotient. The higher the value,

14 In a similar fashion, the 100 most frequent measure nouns occurring with plural kind nouns were listed and queried, resulting in a sample of 871 sentences. As stated in Section 2, the NAC_{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 times NAC_{adj} and 794 times PGC_{adj}). However, the sample is distributed with the data package accompanying this paper.

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.

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

the more often the noun occurs in the $\mathrm{PGC}_{\mathrm{det}}$(proportionally).[16] Additionally, *Kindfreq* and *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. They were added to control for basic frequency effects.

In Section 2.2, it was hypothesised that classes of measure lemmas might have different preferences for the two alternants. 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).

16 It could be argued that some more advanced measure of attraction strength should be used, as is done in Collostructional Analysis (Stefanowitsch & Gries, 2003; Gries & Stefanowitsch, 2004), see also Gries (2015b). Three main points speak against such an approach in the present case. First, the goal here is to quantify how often lemmas occur in the $\mathrm{PGC}_{\mathrm{det}}$ and the $\mathrm{NAC}_{\mathrm{bare}}$, and these constructions do not compete at all but are rather mutually exclusive. Collostructional approaches are not made for such scenarios. Second, the attraction values will be used as regressors in a hierarchical logistic regression and the values resulting from collostructional analysis, i. e., logarithmised Fisher p-values, have a very unfavourable distribution in the case at hand. They cluster around 0 and they include values of $-\infty$. Third, I tried using collexeme strength as a regressor (with smoothing to remedy the mathematical problems), and the results were unsatisfactory compared to the simple quotient used here.

The lemmas that did not fit into either of these classes were labelled *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 NAC_{adj} as hypothesised in Section 2.2.

To capture the influence of 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 pregenerated language model for German. Documents with higher Badness usually contain more incoherent language, shorter sentences, etc. If the PGC_{adj} actually favours higher stylistic levels, a high *Badness* should be correlated with fewer occurrences. 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. A high number of genitives is indicative of higher levels of style. However, the use of this variable as a regressor in the present study might be considered problematic. Since the PGC_{adj} contains a genitive itself, the regressor variable *Genitive* and the document-level variable *Genitives* are not fully independent. Since instances of the PGC_{adj} make up for only a minute fraction of all genitives, I still use *Genitives* as a regressor with the appropriate caveats.

Finally, one variable was 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 PGC_{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' (PGC_{adj}) would be preferred more strongly against *mit einem Stück frischem Brot* (NAC_{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 On statistical analysis

In this section, I justify my choice of statistical models used in Section 3.4. Readers might think that the method used here for modeling grammatical alternations – namely (Hierarchical) Logistic Regression/Generalised Linear (Mixed) Models or GL(M)Ms – does not require much justification. After all, GLMMs have been established as the major tool in the analysis of alternation phenomena. All studies mentioned at the outset of Section 1 use some form of regression/

GLMM. Over the past few years, however, modified or alternative methods have been proposed. While it is impossible to discuss all of these methods here, I just make a few remarks on Bayesian estimation (see Gelman et al., 2014) as it was proposed in Levshina (2016) and 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 (Divjak, 2016, 301–302). 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.

The first and second point cannot be discussed in detail here.[17] With regard to the third point, 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 a strong focus on the rejection of the null hypothesis is characteristic only of Fisher's approach. In the Neyman-Pearson approach, results ideally *favour* the main hypothesis vis-à-vis the alternative hypothesis (cf. Lehmann, 1993, 2011; Perezgonzalez, 2015). Also, especially Neyman-style frequentism has well-known extensions to estimation, for example in the form of confidence intervals (see Greenland et al., 2016, esp. p. 340). 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, given this statement, she does *not* use informative priors and in her footnote 8 (Levshina, 2016, 252) admits that priors were probed using trial and error. So, the proclaimed major advantage of Bayesian modeling was apparently not taken advantage of.[18] Now, Maximum Likelihood Estima-

17 However, there is a number of critical papers by statisticians (even prominent Bayesians) and philosophers of science in which the extreme credibility that has recently been assigned to (subjective and automatic) Bayesian approaches (and especially the standard inductivist interpretation of Bayesianism) is questioned (for example, Gelman & Shalizi, 2013; Mayo, 2011; Senn, 2011).

18 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" and give recommendations how to proceed once posteriors have been obtained from noninformative priors.

tion (MLE) – the traditional method which could have been used instead – is not exactly *frequentist* in the sense of Neyman-Pearson testing theory. MLE, like inductive Bayesianism, conditions on the particular data inasmuch as it searches for the most likely set of parameters given the data. What is more, Bayesian estimators are in fact based on the Likelihood and merely multiply it by the prior (Gelman et al., 2014, 6–8). 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 parameter 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 been obtained with Maximum Likelihood methods. In fact, she admits that changing the priors did not lead to substantially different results in her footnote 8. This is a clear sign that the prior is "swamped by the data" (Freedman, 1999, 1119).

If there had been evidence in Levshina's study that Bayesian and MLE methods did *not* converge, it would have been an occasion to demonstrate the selective superiority of the algorithms used in Bayesian estimation. After all, there are situations where Bayesian estimators can be more robust, namely with heavily censored data, complex hierarchical models, perfect separation, etc. (see Freedman, 1999, Gelman & Hill, 2006, 345–348). I want to state clearly that these points do not in any way invalidate the results presented in Levshina (2016). However, being "Bayesian" is most likely not among its selling points. Additionally, I want to voice the concern that many practitioners are probably already struggling with getting an adequate grasp of advanced statistical methods and that it might therefore be wise to use the more conservative and better understood method if the alternative method is not absolutely required for substantive reasons. In the next section, I estimate the parameters of my hierarchical model with MLE and Markov-Chain Monte Carlo (MCMC) methods (the currently most prominent estimator used in Bayesian settings) to demonstrate their expectable convergence.

3.4 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* (**?**) for Maximum Likelihood Estimation (MLE) and *rstanarm* (Gabry & Goodrich, 2016) for Bayesian Markov-Chain Monte Carlo (MCMC) estimation (see Section 3.3). The purpose is to model the influence of the regressors specified in Table 2 on the probability that the $\mathrm{PGC}_{\mathrm{adj}}$ is

chosen over the $\mathrm{NAC}_{\mathrm{adj}}$. 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 1,134 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 intercept comprises *Cardinal=Yes*, *Measurecase=Nom*, *Kindgender=Masc*, *Measureclass=Physical*, and 0 for all numeric z-transformed regressors. It is -4.328 for the ML estimate and -4.441 for the MCMC estimate.

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 $\mathrm{PGC}_{\mathrm{adj}}$.

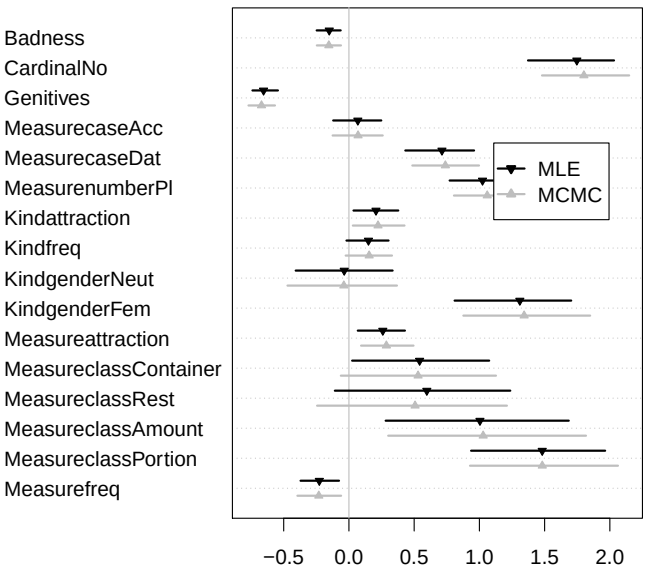

Fig. 1: Coefficients (MLE and MCMC) with 95% confidence intervals (for details see text); the intercept is -4.328 (MLE) and -4.441 (MCMC)

Model level	Regressor	p_{PB}	Factor level	Coefficient		CI low		CI high		CI excludes 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 is -4.328 (MLE) and -4.441 (MCMC)

Standard diagnostics for MLE show that the model quality is quite good. 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 $\text{GVIF}^{1/2\text{df}}$ 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 basic 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 lemma intercepts have standard deviations of $\sigma_{\text{Measurelemma}} = 0.448$ and $\sigma_{\text{Kindlemma}} = 0.604$.

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 excludes 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 inexact Wald approximations and 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). I call the corresponding value p_{PB}, and it is given in the respective columns in Table 3. Only *Kindfreq* ($p_{\text{PB}} = 0.095$) can be seen as slightly too high to be convincing (non-significant).

Finally, I show that MCMC does not necessarily lead to different results. Actually, given the low complexity of the model and the large sample size, it would be surprising if they did (see Section 3.3). 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.[19] The algorithm converged, and for all coefficients, the $\widehat{R}$ diagnostic was exactly 1. The resulting coefficients and intervals as well as an * for "confidence interval does not contain

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

0" 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 given the modest complexity of the model structure and the large sample size. The signs and magnitudes of the coefficients are identical, and the 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.

3.5 Interpretation

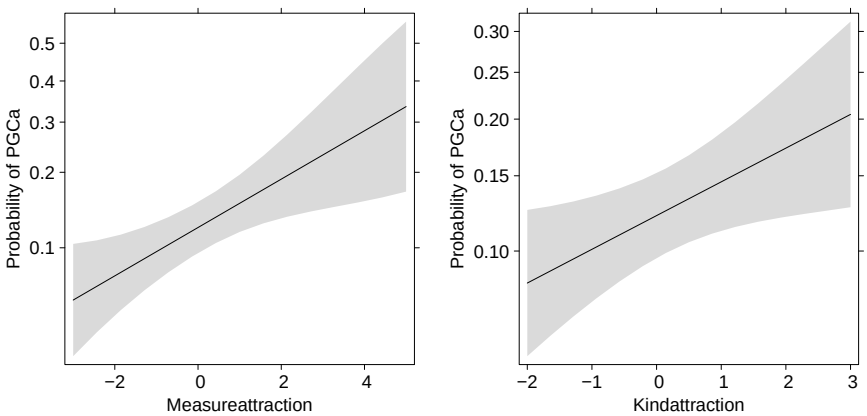

Fig. 2: Effect plots for the regressors *Measureattraction* and *Kindattraction*; y-axes are not aligned

The results reported in Section 3.4 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).[20] The effect is as expected: if a lemma appears relatively more often in the $\mathrm{PGC}_{\mathrm{det}}$

20 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. The vertical bars (categorical variables), and the grey areas (continuous variables) are asymptotic 95% confidence intervals calculated from *glmer*. They are not bootstrapped. Readers should be aware that the axes are specifically scaled so as to result in a linear plot, and that the range of the axes varies between plots.

(compared to its frequency in the $\mathrm{NAC}_{\mathrm{bare}}$), the $\mathrm{PGC}_{\mathrm{adj}}$ tends to be 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.

An interesting picture emerges for the lemma frequencies. A higher-than-average lemma frequency of measure nouns favours the $\mathrm{NAC}_{\mathrm{adj}}$ ($\beta_{\mathrm{Measurefreq}} = -0.231$, $p_{\mathrm{PB}} = 0.005$), which is 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 $\mathrm{PGC}_{\mathrm{adj}}$($\beta_{\mathrm{Kindfreq}} = 0.146$, $p_{\mathrm{PB}} = 0.095$). 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.

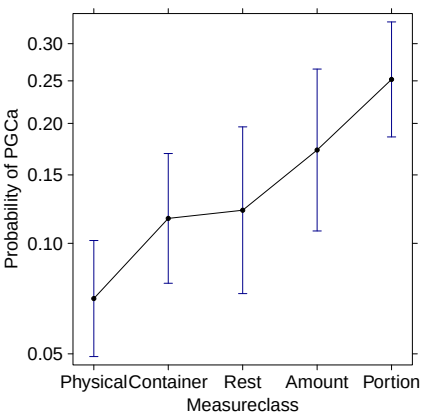

Fig. 3: Effect plot for the regressor *Measureclass*

In Section 2.2, it was also hypothesised that classes of measure nouns with a higher degree of grammaticalisation should favour the $\mathrm{NAC}_{\mathrm{adj}}$. The *Measureclass* second-level predictor was successfully estimated ($p_{\mathrm{PB}} = 0.001$). 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 $\mathrm{NAC}_{\mathrm{adj}}$. At the other end of the scale, nouns denoting natural portions like *Haufen* 'heap', *Bündel* 'bundle', *Schluck* 'gulp' favour the $\mathrm{PGC}_{\mathrm{adj}}$. These are referential nouns, confirming the hypothesis that it is prototypical of the PGC to contain two referential nouns, while the NAC prototypically only contains one (the kind noun).

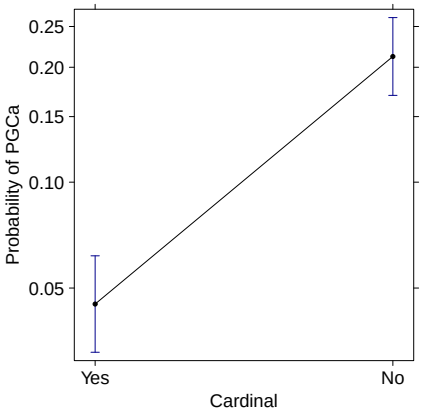

Fig. 4: Effect plot for the regressor *Cardinal*

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 alternant ($p_{\mathrm{PB}} = 0.001$), 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 style-related proxy variables point to the expected direction. Increased *Badness* of the document favours the $\mathrm{NAC}_{\mathrm{adj}}$ ($\beta_{\mathrm{Badness}} = -0.152$, $p_{\mathrm{PB}} = 0.002$), and so does a lower density of genitives ($\beta = -0.693$, $p_{\mathrm{PB}} = 0.001$). While these are merely proxies to style (and partially circular in the case of *Genitives*), this result can at least encourage future work into stylistic 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 $\beta_{\mathrm{MeasurecaseDat}} = 0.705$ (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.

4 Experiments

4.1 Experiment 1: forced-choice

In the two experiments reported in this section, I use probabilities for the alternating constructions calculated for attested material, and I correlate these probabilities with the participants' reactions. Thus, a direct link can be established between output material found in corpora and the behaviour of linguistic agents (see also Section 1). 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 alternants in these sentences are used as the main regressor in both studies.

The first experiment tests preferences for constructions explicitly. Ford & Bresnan (2013) use the *split-100* task in which participants have to distribute 100 points between the alternatives, assigning more points to more natural sounding alternative. In essence, participants distribute a probability mass between two alternants, which is intended to produce more subtle results compared to a two-alternative forced-choice task such as in Rosenbach (2013), where participants have to choose one of two alternants. The split-100 paradigm has been criticised in Arppe & Järvikivi (2007). The criticism was reiterated in Divjak et al. (2016), where they use a forced-choice task. In Verhoeven & Temme (2017, to appear), it was shown that results from forced-choice and split-100 experiments mostly converge (with some numeric intricacies related to the non-linear distribution of preferences for alternants). I present a forced-choice experiment in this Section and not a split-100 experiment, mainly because in a dry run of the experiment, participants complained about the unnaturalness of distributing a probability mass across two alternants and tended to produce ratings of 0 and 100 (and to a lesser degree 50). Participants had to choose between two sentences that differed only in that one contained the $\mathrm{NAC}_{\mathrm{adj}}$ and the other one contained the $\mathrm{PGC}_{\mathrm{adj}}$. The analysis compares the probabilities assigned to the stimuli by the corpus-derived model with the frequency with which participants chose the alternants for the same stimuli.

There were 24 participants (native speakers of German without reading or writing disabilities) aged 19 to 30 living permanently in Berlin, who were recruited from introductory linguistics courses at Freie Universität Berlin. 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

	Masculine/Neuter	Feminine
high prob. for PGC_{adj}	4 sentences	4 sentences
low prob. for PGC_{adj}	4 sentences	4 sentences

Table 4: The four groups of sentences chosen as stimuli; in each group of four sentences, combinations of important factor values were made unique whenever possible

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 factors as much as possible. The approach is also justified because according to the theoretical assessment in Section 2.2, the choice of alternants depends mostly on a very local constructional context. I sampled 16 MNPs from the concordance and 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 did the choice of lexical material within the MNP. Eight sentences contained masculine or neuter kind nouns, and the other eight contained feminine kind nouns. Furthermore, in each of the masculine/neuter and feminine groups, four sentences originally containing the NAC_{adj} and four sentences originally containing the PGC_{adj} were chosen. More precisely, the sentences were sampled as *highly prototypical examples* of PGC_{adj} (high probability assigned by GLMM) and NAC_{adj} (low probability assigned by GLMM), respectively.[21] High and low probabilities 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 final pairs of stimuli were the sentence containing the attested and preferred alternant (according to the corpus GLMM) on the one hand and a modified version containing the dispreferred alternant on the other hand. They were presented next to each other, and a 20 second time limit for each choice was set.[22] The position on the screen (left/right) and the order of sentences were randomised for each participant. As fillers, 23 pairs of sentences exemplifying similar but unrelated alternation phenomena from German morpho-syntax were used. Thus, participants saw 39 pairs of sentences and 78 sentences in total.

21 Remember from Section 3 that the model predicts the probability that the PGC_{adj} is chosen over the NAC_{adj}.

22 No participant ever exceeded the time limit.

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).

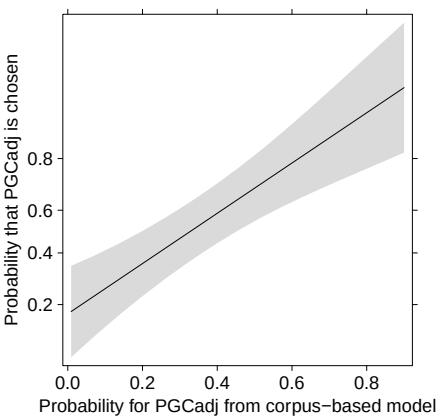

Fig. 5: Effect plot for the multilevel logistic regression in the forced-choice experiment: predictability of participants' choices using the probabilities derived from the corpus-based GLMM

Then, a multilevel logistic regression was specified with the probability of the $\mathrm{PGC}_{\mathrm{adj}}$ predicted for each sentence by the corpus-based GLMM as the only fixed effect *Modelprediction*.[23] A random intercept and slope were added for the individual sentence (item) in order to catch idiosyncrasies of single sentences. Also, a random intercept and slope for participants was added.[24] Coefficients were estimated with Maximum Likelihood Estimation (*lmer* function from *lme4*). The number of observations was n=384.

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

24 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.

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).[25] 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 ends, one participant chose the PGC_{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. The 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 PGC_{adj} predicted from usage data, the more often participants chose the PGC_{adj} alternant in the forced-choice task. In summary, the forced-choice experiment clearly succeeded in corroborating 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

The second experiment tests preferences more implicitly. It is expected that reading less prototypical alternants (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). In this section, a self-paced reading experiment is therefore presented. In a very similar fashion, Divjak et al. (2016) apply the self-paced reading paradigm in the validation of corpus-based models. The analysis compares the corpus-derived probabilities with potential lags in reading time for sentences with the preferred and the non-preferred constructions.

Concretely, the exact same stimuli as in the forced-choice experiments were used. Each participant read both the 16 sentences with the alternant predicted by the corpus model and the 16 modified sentences with the alternant that the

25 I use σ^f_r to denote the standard deviation of the random intercepts for the fixed effect f varying by random effect r.

corpus model did not predict.[26] 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. 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, as was the order within each block. The sentences from each pair of alternants 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 between the first and the second blocks. There were 38 participants recruited in exactly the same manner as for the experiment reported in Section 4.1. None of them had ever participated in any kind of reading experiment, and none of them took part in the first experiment. 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, for instance 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 residualised reading 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.

The probabilities derived from the corpus GLMM (*Modelprediction*) were added as the main regressor of interest. It should be remembered that the corpus GLMM predicts the probability of the $\mathrm{PGC}_{\mathrm{adj}}$. As a consequence, the higher the GLMM prediction is, the more prototypical the sentence is for containing the $\mathrm{PGC}_{\mathrm{adj}}$. It is therefore expected that reading times are higher when the value of *Modelprediction* is higher but the sentence contains the $\mathrm{NAC}_{\mathrm{adj}}$. However, when the sentence contains the $\mathrm{PGC}_{\mathrm{adj}}$, reading times should be lower when *Model-*

26 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	CI excludes 0
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

prediction is higher. To account for this, an interaction between *Modelprediction* 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 the usual increase in reading speed during an experiment run. Random intercepts were specified for *Participant* and *Item* (the 16 sentence pairs are one *Item* each).[27]

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 $p_{\text{PB}} = 0.036$. The pseudo-coefficients of determination are $R^2_m = 0.237$ and $R^2_c = 0.346$. There is thus some inter-subject variation and a noteworthy correlation between the reactions of the subjects and the corpus-derived probabilities. The fact that the correlation is not stronger can be explained in parts

The effect plot for the effect of interest is shown in Figure 6. The estimate for the sentences with NAC_{adj} is obviously imprecise and no differences in reading times are observed. There is a clearer effect in the sentences with PGC_{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 PGC_{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 in the right panel of Figure 6. This fits into the general picture inasmuch

27 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.

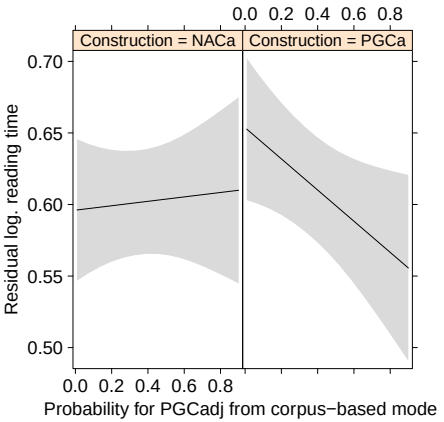

Fig. 6: Effect plot for the LMM in the self-paced reading experiment: modeling participants' residualised log reading times on the probabilities given by the corpus-based GLMM

as the construction with the lower frequency might be developing towards a more sharply defined prototype.[28] Conversely, the NAC_{adj} (like the NAC in general) might be the highly frequent default which does not incur a reading time penalty, even if it is not the optimal choice in the given context and with the given lexical material.

This concludes the report of the two experiments. In the final section, I take stock and summarise the contribution of the present study to the research on alternations in cognitive linguistics.

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

5 Conclusions

This paper stands in a now ten year-old tradition of research on grammatical alternations using corpus and experimental data. In my view, the main tenets of this line of research are: (i) Language, viewed from a cognitively realistic angle, is a probabilistic phenomenon and cannot be modelled appropriately within Aristotelian frameworks that assume discrete categories. (ii) Corpora are collections of usage events (language production) and can therefore be used to evaluate both the claim made in (i) and specific theoretical claims (in case studies) about factors influencing speakers' decisions to use specific forms or constructions. (iii) Given (ii), we expect results from corpus analyses and from appropriate experiments to yield similar results, not necessarily as a form of validation of the corpus-based findings, but as converging evidence. I consider these three points to be of utmost importance because they clearly set this approach to linguistic research apart from *both* Aristotelian frameworks *and* introspective, non-empirical, and anti-quantitative versions of cognitive linguistics (see Dąbrowska, 2016 for a pithy philippic against such approaches).

The present paper adds to the evidence that all of the aforementioned three points are correct. A grammatical alternation in German measure NPs was examined using corpus data based on factors partly derived from existing accounts, formulated in terms of construction prototypes. The preferences extracted from the DECOW web corpus were confirmed in a forced-choice experiment, in which participants explicitly chose alternants in line with the probabilities derived from the corpus-based model. In a more implicit self-paced reading experiment, it was shown that the much rarer alternant brings about a reading time penalty except in cases for which the corpus model predicts very high probability for this alternant.

Future work could extend these results and provide a general picture of the constructions expressing measurements (see Section 2.1). This would be a much more complicated task given that the choices then would no longer be binary and that the meaning of the alternative ways of expressing measurements are semantically more varied. Finally, I want to point out that German is mildly under-researched in the specific framework used here. This is quite surprising given the fact that German morpho-syntax is famous for its alternations, which are usually called *Zweifelsfälle* ('cases of doubt') in the traditional literature (Duden, 2011; Klein, 2009). Instead of being drowned in normative, descriptive, or didactic discussions, they could serve as ideal test cases in cognitive linguistics.

Acknowledgment: I thank (in alphabetical order) Felix Bildhauer, Susanne Flach, Elizabeth Pankratz, Samuel Reichert, Ulrike Sayatz, and Christian Zimmer for valuable discussions and comments. Also, I would like to thank two reviewers for Cognitive Linguistics with deep linguistics knowledge and knowledge of German grammar for comments which helped to improve the paper a lot. I thank Dagmar Divjak who, in her role as editor for Cognitive Linguistics, also read the paper thoroughly and provided very useful comments.

Furthermore, I thank Ulrike Sayatz for helping me to recruit the participants for the experiments. Elizabeth Pankratz thankfully also fixed my English. Finally, I am grateful to my student assistants Kim Maser for her work on the annotation of the concordances and Luise Rißmann for supervising most of the experiments. The research presented here was made possible in part through funding from the *Deutsche Forschungsgemeinschaft* (DFG, personal grant SCHA1916/1-1).

References

Arppe, Antti & Juhani Järvikivi. 2007. Every method counts: combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159. 10.1515/cllt.2007.009.

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. 10.1080/23273798.2015.1065336.

Barker, Chris. 1998. Partitives, double genitives and anti-uniqueness. *Natural Language and Linguistic Theory* 16(4). 679–717. 10.1111/1754-9485.12268.

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. 10.1016/j.jml.2012.11.001.

Barsalou, Lawrence W. 1990. On the indistinguishability of exemplar memory and abstraction in category representation. In Thomas K. Srull & Robert S. Wyer (eds.), *Advances in social cognition, volume III: Content and process specificity in the effects of prior experiences*, 61–88. Hillsdale: Lawrence Erlbaum Associates. 10.1017/s0140525x00051591.

Bhatt, Christa. 1990. *Die syntaktische Struktur der Nominalphrase im Deutschen*. Tübingen: Narr. 10.1515/zfgl.2003.31.3.327.

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. 10.2200/s00508ed1v01y201305hlt022.

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. 10.18653/v1/w16-2601.

Brems, Lieselotte. 2003. Measure noun construction: An instance of semantically-driven grammaticlization. *International Jounal of Corpus Linguistics* 8(2). 283–312. 10.1075/ijcl.8.2.05bre.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base* (Studies in Generative Grammar), 77–96. Berlin/New York: De Gruyter Mouton. 10.1515/cllt.2011.011.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In Gerlof Bouma, Irene Krämer & Joost Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Koninklijke Nederlandse Akadmie van Wetenschappen.

Bresnan, Joan & Marilyn Ford. 2010. Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86(1). 168–213. 10.1353/lan.0.0189.

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. 10.1016/j.lingua.2007.02.007.

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. 10.18637/jss.v076.i01.

Conaway, Nolan & Kenneth J. Kurtz. 2016. Similar to the category, but not the exemplars: a study of generalization. *Psychonomic Bulletin & Review* 24. 1312–1323. DOI 10.3758/s13423-016-1208-1.

Dąbrowska, Ewa. 2016. Cognitive linguistics' seven deadly sins. *Cognitive Linguistics* 27(4). 479–491. 10.1515/cog-2016-0059.

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. 10.1016/j.langsci.2015.05.007.

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. 10.1515/9783110435597-017.

Divjak, Dagmar & Antti Arppe. 2013. Extracting prototypes from exemplars what can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). 221–274. 10.1515/cog-2013-0008.

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. 10.1037/xlm0000410.

Divjak, Dagmar & Stefan Th. Gries. 2008. Clusters in the mind? Converging evidence from near-synonymy in Russian. *The Mental Lexicon* 3(2). 188–213. 10.1075/ml.3.2.03div.

Duden. 2011. *Richtiges und gutes Deutsch - Das Wörterbuch der sprachlichen Zweifelsfälle*. Mannheim/Zürich: Dudenverlag 7th edn. 10.1002/ange.19330465002.

Eisenberg, Peter. 2013. *Grundriss der deutschen Grammatik: Der Satz*. Stuttgart: Metzler 4th edn. 10.1007/978-3-476-03762-6.

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. 10.1007/978-3-663-05399-6_9.

Ford, Marilyn & Joan Bresnan. 2013. Using convergent evidence from psycholinguistics and usage. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 295–312. Cambridge, MA: Cambridge University Press. 10.1017/cbo9780511792519.020.

Fox, John. 2003. Effect displays in R for Generalised Linear Models. *Journal of Statistical Software* 8(15). 1–27. 10.2307/271037.

Fox, John & Georges Monette. 1992. Generalized collinearity diagnostics. *Journal of the American Statistics Association* 87. 178–183. 10.2307/2290467.

Freedman, David. 1999. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics* 27(4). 1119–1140. 10.1214/aos/1017938917.

Gabry, Jonah & Ben Goodrich. 2016. *rstanarm: Bayesian applied regression modeling via stan*. 10.1007/s11222-016-9709-3. R package version 2.12.1.

Gallmann, Peter & Thomas Lindauer. 1994. Funktionale Kategorien in Nominalphrasen. *Beiträge zur Geschichte der deutschen Sprache* 116. 10.1515/bgsl.1994.116.1.1.

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. 10.2307/2965436.

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press. 10.1017/cbo9780511790942.

Gelman, Andrew & Cosma Rohilla Shalizi. 2013. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* 66. 8–38. 10.1111/j.2044-8317.2011.02037.x.

Gerstenberger, Laura. 2015. Number marking in German measure phrases and the structure of pseudo-partitives. *Journal of Comparative Germanic Linguistics* 18. 93–138. 10.1007/s10828-015-9074-1.

Goethem, Kristel Van & Philippe Hiligsmann. 2014. When two paths converge: Debonding and clipping of Dutch 'reuze'. *Journal of Germanic Linguistics* 26(1). 31–64. 10.1017/s1470542713000172.

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. 10.1017/s1470542715000112.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman & Douglas G. Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31. 337–350. 10.1007/s10654-016-0149-3.

Gries, Stefan Th. 2003. Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1. 1–27. 10.1075/arcl.1.02gri.

Gries, Stefan Th. 2015a. The most underused statistical method in corpus linguistics: Multilevel (and mixed-effects) models. *Corpora* 10(1). 95–126. 10.3366/cor.2015.0068.

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. 10.1515/9783110435597-018.

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. 10.1515/cllt.2005.1.1.1.

Griffiths, Thomas L., Kevin R. Canini, Adam N. Sanborn & Daniel J. Navarro. 2009. Unifying rational models of categorization via the hierarchical dirichlet process. In *Proceedings of the 29th annual conference of the cognitive science society*, 323–328. Mahwah: Erlbaum.

Hahn, Ulrike, Mercè Prat-Sala, Emmanuel M. Pothos & Duncan P. Brumby. 2010. Exemplar similarity and rule application. *Cognition* 114(1). 1–18. 10.1016/j.cognition.2009.08.011.

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. 10.18637/jss.v059.i09.

Hankamer, Jorge & Line Mikkelsen. 2008. Definiteness marking and the structure of Danish pseudopartitives. *Journal of Linguistics* 44(2). 317–346. 10.1017/s0022226708005148.

Hentschel, Elke. 1993. Flexionsverfall im Deutschen? Die Kasusmarkierung bei partitiven Genitiv-Attributen. *Zeitschrift für Germanistische Linguistik* 21(3). 320–333. 10.1515/9783111566658.17.

Hintzman, Douglas L. 1986. Schema abstraction in a multiple-trace memory model. *Psychological Review* 93(4). 411–428. 10.1037//0033-295x.93.4.411.

Kaiser, Elsi. 2013. Experimental paradigms in psycholinguistics. In Robert J. Podesva & Devyani Sharma (eds.), *Research methods in linguistics*, 135–168. Cambridge University Press. 10.3765/salt.v25i0.3436.

Kilgarriff, Adam. 2006. Googleology is bad science. *Computational Linguistics* 33(1). 147–151. 10.1162/coli.2007.33.1.147.

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. 10.1007/s40607-014-0009-9.

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. 10.1007/bf00316024.

Köpcke, Klaus-Michael. 1995. Die Klassifikation der schwachen Maskulina in der deutschen Gegenwartssprache – Ein Beispiel für die Leistungsfähigkeit der Prototypentheorie. *Zeitschrift für Sprachwissenschaft* 14(2). 159–180. 10.1515/zfsw.1995.14.2.159.

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. 10.1075/slcs.55.11kop.

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). 10.1075/lis.28.08bel.

Labov, William. 1966. *The social stratification of English in New York City*. Washington, DC: Center for Applied Linguistics. 10.1525/aa.1964.66.suppl_3.02a00120.

Labov, William. 1969. Contraction, deletion, and inherent variability of the english copula. *Language* 45(4). 715–762. 10.2307/412333.

Lee, Michael D. & Wolf Vanpaemel. 2008. Exemplars, prototypes, similarities, and rules in category representation: an example of hierarchical Bayesian analysis. *Cognitive Science* 32. 1403–1424. 10.1080/03640210802073697.

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. 10.1007/978-1-4614-1412-4_19.

Lehmann, Erich L. 2011. *Fisher, Neyman, and the creation of classical statistics*. New York, NY: Springer. 10.1007/978-1-4419-9500-1.

Levshina, Natalia. 2016. When variables align: A Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2). 235–268. 10.1515/cog-2015-0054.

Löbel, Elisabeth. 1986. Apposition in der Quantifizierung. In Armin Burkhardt & Karl-Hermann Körner (eds.), *Pragmantax. Akten des 20. Linguistischen Kolloquiums Braunschweig 1985*, 47–59. Tübingen: Niemeyer. 10.1164/rccm.2309009.

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. 10.1075/la.6.10lob.

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. 2011. How can we cultivate Senn's ability? *Rationality, Markets, and Morals* 3. 14–18. 10.1097/00004045-199503000-00012.

Medin, Douglas L. & Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review* 85(3). 207–238. 10.1037//0033-295x.85.3.207.

Minda, John Paul & J. David Smith. 2001. Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(3). 775–799. 10.1037/0278-7393.27.3.775.

Minda, John Paul & J. David Smith. 2002. Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28(2). 275–292. 10.1037/0278-7393.28.2.275.

Müller, Sonja. 2014. Zur Anordnung der Modalpartikeln "ja" und "doch": (In)stabile Kontexte und (non)kanonische Assertionen. *Linguistische Berichte* 238. 165–208. 10.1007/bf00386549.

Murphy, Gregory L. 2003. Ecological validity and the study of concepts. In Brian H. Ross (ed.), *Psychology of learning and motivation - advances in research and theory*, 1–41. New York: Elsevier. 10.1016/S0079-7421(03)01010-7.

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. 10.1111/j.2041-210x.2012.00261.x.

Nesset, Tore & Laura A. Janda. 2010. Paradigm structure: evidence from Russian suffix shift. *Cognitive Linguistics* 21(4). 699–725. 10.1515/cogl.2010.022.

Peirce, Jonathan W. 2007. Psychopy – Psychophysics software in Python. *Journal of Neuroscience Methods* 162(1–2). 8–13. 10.1016/j.jneumeth.2006.11.017.

Perezgonzalez, Jose D. 2015. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology* 6(223). 1–11. 10.3389/fpsyg.2015.00223.

Posner, Michael I. & Steven W. Keele. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology* 77(3). 353–363. 10.1037/h0025953.

R Core Team. 2014. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. 10.1111/j.1467-8624.2009.01290.x.

Ramscar, Michael & Robert F. Port. 2016. How spoken languages work in the absence of an inventory of discrete units. *Language Sciences* 53. 58–74. 10.1016/j.langsci.2015.08.002.

Rosch, Eleanor. 1978. Principles of categorization. In Eleanor Rosch & Barbara B. Lloyd (eds.), *Cognition and categorization*, 27–48. Hillsdale: Lawrence Erlbaum Associates. 10.1016/b978-1-4832-1446-7.50028-5.

Rosenbach, Anette. 2013. Combining elicitation data with corpus data. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 278–294. Cambridge, MA: Cambridge University Press. 10.1017/cbo9780511792519.019.

Rosseel, Yves. 2002. Mixture models of categorization. *Journal of Mathematical Psychology* 46(2). 178–210.

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. 10.1111/j.1467-7687.2010.01001.x.

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. 10.1075/la.6.08sch.

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. 10.1007/s00294-003-0404-5.

Schäfer, Roland. 2016. Prototype-driven alternations: the case of German weak nouns. *Corpus Linguistics and Linguistic Theory* ahead of print. 10.1515/cllt-2015-0051.

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. 10.3115/v1/w14-0402.

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. 10.2200/s00508ed1v01y201305hlt022.

Schäfer, Roland & Ulrike Sayatz. 2014. Die kurzformen des indefinitartikels im deutschen. *Zeitschrift für Sprachwissenschaft* 33(3). 215–250. 10.1515/zfs-2014-0008.

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. 10.1075/wll.19.2.04sch.

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. 10.2307/413978.

Senn, Stepen J. 2011. You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets, and Morals* 2. 48–66. 10.1002/ss.20167.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243. 10.1075/ijcl.8.2.03ste.

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. 10.7557/12.128.

Storms, Gert, Paul De Boeck & Wim Ruts. 2000. Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language* 42. 51–73. 10.1006/jmla.1999.2669.

Tagliamonte, Sali A. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Malden/Oxford: Wiley-Blackwell. 10.1002/9781118455494.

Tummers, José, Kris Heylen & Dirk Geeraerts. 2005. Usage-based approaches in cognitive linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2). 225–261. 10.1515/cllt.2005.1.2.225.

Vanpaemel, Wolf. 2016. Prototypes, exemplars and the response scaling parameter: a Bayes factor perspective. *Journal of Mathematical Psychology* 72. 183–190. 10.1016/j.jmp.2015.10.006.

Vanpaemel, Wolf & Gert Storms. 2008. In search of abstraction: The varying abstraction model of categorization. *Psychonomic Bulletin & Review* 15(4). 732–749. 10.3758/PBR.15.4.732.

Verbeemen, Timothy, Wolf Vanpaemel, Sven Pattyn, Gert Storms & Tom Verguts. 2007. Beyond exemplars and prototypes as memory representations of natural concepts: a clustering approach. *Journal of Memory and Language* (537–554). 10.1016/j.jml.2006.09.006.

Verhoeven, Elisabeth & Anne Temme. 2017, to appear. Word order acceptability and word order choice. In Sam Featherston, Robin Hörnig, Reinhild Steinberg, Birgit Umbreit & Jennifer Wallis (eds.), *Linguistic Evidence 2016 Online Proceedings*, Tübingen: Universität Tübingen. 10.1515/ling-2016-0018.

Voorspoels, Wouter, Wolf Vanpaemel & Gert Storms. 2011. A formal ideal-based account of typicality. *Psychonomic Bulletin & Review* 18. 1006–1014.

Vos, Riet. 1999. *A grammar of partitive constructions* (Tilburg dissertation in language studies). Tilburg: Tilburg University. 10.2307/2919069.

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. 10.1007/s10579-016-9343-x.

Zifonun, Gisela, Ludger Hoffmann & Bruno Strecker. 1997. *Grammatik der deutschen Sprache*, vol. 3. 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. 10.1515/bgsl-2015-0001.

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. 10.1111/j.2041-210x.2009.00001.x.