Roland Schäfer*

Competing Constructions for German Measure Noun Phrases: from Usage Data to Experiment

Abstract: In this paper, a case alternation occurring in German measure noun phrases is examined from the perspective of cognitively oriented corpus linguistics. In pseudo-partitive constructions such as *eine Tasse guter Kaffe* 'a cup of good coffee', the embedded kind noun either agrees in its case with the measure head noun or it is a genitive as in *eine Tasse guten Kaffees*. This alternation occurs only under highly specific syntactic conditions. The analysis is formulated in terms of Prototype Theory. I assume that the variants are modelled after non-alternating neighbouring constructions, which represent the prototypes more clearly. Primarily, it is argued that the frequencies with which individual lemmas appear in the prototypical constructions partially predicts which variant is chosen in the alternating case, and that the genitive construction is prototypical of more strongly grammaticalised measure nouns. I present a large-scale corpus study using the DECOW web corpus which supports the theoretical analysis. In the statistical analysis of the corpus study, it is discussed whether Bayesian estimators are intrinsically better than Likelihood Maximisation methods, and all results are estimated with Maximum Likelihood and Bayesian methods for comparison. Finally, a forced choice experiment and a self-paced reading experiment are reported. In both experiments, the usage-based corpus analysis is successfully correlated with the behaviour of cognitive agents. The present study therefore contributes to a now well-established area of research into alternations using both corpus and experimental methods.

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

1 Cognitively oriented corpus linguistics

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

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

available with no clear difference in acceptability, function, or meaning. The study of alternations has a long history in cognitively oriented corpus linguistics (Bresnan et al., 2007; Bresnan & Hay, 2008; Bresnan & Ford, 2010; Divjak & Arppe, 2013; Gries, 2015a; Nesset & Janda, 2010, to name just a few publications). This area of research is based on the assumption that language is a probabilistic phenomenon (Bresnan, 2007) where choices of variants 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 variant is chosen.

For such probabilistic generalisations estimated on corpus data, there has always been an interest in correlating the findings 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 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*. This position, which I adopt, does not, however, imply that we can in some way *deduce mental representations from patterns of use*, i. e., from corpus data, as Dąbrowska (2016, 486–487) convincingly argues.[1] It would be highly surprising if this were possible. 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. Under the standard approach, we pre-specify a theory of cognitive representation. Then, predictions are derived from this theory *before* the experiment or the corpus study is conducted to *test* the theory. Technical problems of statistical inference aside (see also Section 3.3), a successful experiment corroborates the theory, and a failed experiment weakens the theory. The simplest assumption would be that the same approach is adequate for corpus studies.

The present study is conducted within the general paradigm of cognitive corpus linguistics and includes a comparison of the corpus findings with results from two experiments. As a theoretical framework, I assume a model of similarity-based classification in the form of Prototype Theory. Protoype Theory has been

1 In its colloquial meaning, the verb *deduce* seems appropriate here. However, in terms of the scientific method, we are most likely talking about *induction* rather than *deduction*.

used in alternation research in cognitively oriented corpus linguistics, see Divjak & Arppe (2013); Gries (2003). It is assumed that the variables annotated in a corpus study on an alternation phenomenon define prototypes for the alternating variants. The similarity of a given feature vector (in a concrete sentence where one of the variants is used) determines to a large extent which variant is chosen by speakers. This entails that a variant of Prototype Theory with *features* is assumed (Rosch, 1978), instead of the monolithic prototypes of earlier versions. While Prototype Theory is well suited for modeling constructional choices, 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 experimental settings). Prototype Theory and Exemplar Theory differ mainly in whether they assume higher-level abstractions as part of the cognitive representation (Prototype Theory) or not (Exemplar Theory). I am sceptical that corpus analysis alone could ever decide which theory is more adequate, given that substantial doubt has been voiced whether even experimental methods are ultimately able to do so (Barsalou, 1990).[2] Prototype Theory (with feature abstractions) is preferred here simply because it fits the established alternation modeling paradigm, which relies on features being weighted in statistical models. In Section 2.2, a prototype-theoretical parlance is therefore adopted.

In the remainder of the paper, I introduce the alternation phenomenon (Section 2.1) and suggest a theory-driven set of factors influencing the alternation (Section 2.2). Then, the corpus study is presented, including an appropriate statistical analysis (Section 3). Two experiments that correlate results from the corpus study with experimental results are then reported (Section 4) before I sum up the paper in Section 5.

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.

2 In Divjak & Arppe (2013), it was shown using corpus data how both models form a converging picture.

I use the term *measure noun phrase* (MNP) to refer a noun phrase (NP) in which a kind-denoting (count or mass) noun depends on another noun that specifies a quantity of the objects or the substance denoted by the kind-denoting noun. I call the kind-denoting noun simply 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*). In a similar vein, Brems (2003, 284) considers such nouns denoting containers etc. as measure nouns "which, strictly speaking, do not designate a 'measure', but display a more nebulous potential for quantification" (see also Koptjevskaja-Tamm, 2001, 530, and Rutkowski, 2007, 338).

In the case at hand, three different syntactic configurations need to be distinguished w. r. t. case assignment inside German MNPs. 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).[3] Here, the kind noun is in the genitive, and I refer to the construction in (1) as the *Pseudo-partitive Genitive Construction* (PGC).

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

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

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

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

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

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

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

This construction is usually classified as a *Narrow Apposition Construction* (Löbel, 1986), henceforth NAC.[4] Notice that the unavailability of the genitive on the kind noun can be seen as following from a rather quirky constraint that genitive NPs in German require the presence of some strongly case-marked element (determiner or adjective) in addition to the head noun in order to be acceptable (Gallmann & Lindauer, 1994; Schachtl, 1989; see also Eisenberg, 2013, 160).

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

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

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

The distribution of the case patterns in the NAC (case identity between the measure noun and the kind noun) and in the PGC (genitive on the kind noun,

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

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

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

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

i. alleged alternatives to the $\mathrm{NAC}_{\mathrm{adj}}$ and the $\mathrm{PGC}_{\mathrm{adj}}$,
ii. similar constructions with plural/collective kind nouns,
iii. grammaticalised non-inflected measure nouns,
iv. alternative constructions for expressing quantities.

First of all, (i) refers to claims found in some grammars that a generic nominative, accusative, and even dative on the kind noun are used instead of the genitive ($\mathrm{PGC}_{\mathrm{adj}}$) or case agreement ($\mathrm{NAC}_{\mathrm{adj}}$). Overviews can be found in Hentschel, 1993 and Zimmer, 2015. Sentence (4) shows a putative generic nominative on the kind noun inside an accusative MNP.

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

It was shown empirically in Hentschel (1993) that such variants are de facto not acceptable. Also, in my corpus sample, they simply did not occur. Even if they are accepted by some speakers, their extremely low frequency makes it virtually impossible to study them using corpus linguistic methods (Section 3), and I consequently ignore them.

Turning to (ii), readers might have noticed that so far only MNPs denoting quantities of substances (mass kind nouns such as *ein Glas roter Wein/roten Weines* 'a glass of red wine') have been discussed. If the kind noun is a plural count noun as in *ein Sack kleine Äpfel* ($\mathrm{NAC}_{\mathrm{adj}}$) or *ein Sack kleiner Äpfel* ($\mathrm{PGC}_{\mathrm{adj}}$) '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.[6] While this will play a role in the interpretation of

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

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.

As for (iii), some measure nouns have been grammaticalised in a way that they always appear non-inflected. They are typical measure nouns like *Gramm* 'gram', *Pfund* 'pound' or *Prozent* 'percent', which have no normal plural forms at all.[7] I treat these cases like other measure nouns because they enter into both the NAC_{adj} as in (5a) and the PGC_{adj} as in (5b). In Section 2.2, however, degrees of grammaticalisation as a factor influencing the alternation will be discussed.

(5) a. zwei Gramm brauner Zucker

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

Finally, (iv) suggests that there are alternative ways of expressing similar quantificational meanings. In the variationist tradition, which is strongly influenced by Labovian sociolinguistics, the *principle of accountability* would dictate that proper studies should examine a variationist *variable*, i. e., all different *ways of saying the same thing* (Labov, 1966, Labov, 1969, for an overview see Tagliamonte, 2012). In the case at hand, the variable might be something like *measuring quantities of substances and collections*. I argue that it is fully justified to focus narrowly on 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. First of all, the analytic (pseudo-)partitive with *von* 'of' is only available as an alternative to the PGC if the kind noun phrase contains a (definite or indefinite) determiner as in (6).

(6) 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.

7 Plurals *Pfunde* and *Prozente* have special meanings and have very restricted uses, mostly in idiomatic expressions and prefabs. They cannot be used in MNPs.

Second, constructions with *voll* (*von*)/*voller* 'full (of)' and *mit* 'with' are available, see (7).

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

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

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

These construction have very idiosyncratic properties (the construction with *voller* is discussed in Zeldes, to appear. The construction with *mit* is discussed in Bhatt, 1990), they are not semantically equivalent to the NAC and the PGC, and they are only available for a small subset of measure nouns. All of these constructions are used only with measure nouns denoting containers, and the whole NP always refers to the container, and never to the quantity contained in it. They are consequently incompatible with measure nouns denoting natural portions such as *Schluck* 'gulp' or *Haufen* 'heap' and strongly grammaticalised nouns such as *Gramm* 'gram'. This disqualifies them as alternatives to the $\mathrm{NAC}_{\mathrm{adj}}$ or $\mathrm{PGC}_{\mathrm{adj}}$ in most contexts, and they are thus decidedly *not* more ways of saying the same thing. It is therefore reasonable to focus on the two well-defined variants.

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

2.2 Factors controlling the alternation

This section briefly reviews existing analyses of the $\mathrm{PGC}_{\mathrm{adj}}$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. I discuss four main aspects: first, the different syntactic structures of the variants $\mathrm{NAC}_{\mathrm{adj}}$ and $\mathrm{PGC}_{\mathrm{adj}}$ in relation to their syntactic prototypes $\mathrm{NAC}_{\mathrm{bare}}$ and $\mathrm{PGC}_{\mathrm{det}}$; second,

degrees of grammaticalisation of the measure noun associated with the two prototypes; third, a semantic prototypicality effect of the degrees of grammaticalisation (a preference to occur with cardinals); fourth, preferences in different registers.

My analysis is based on the idea that in the relevant syntactic structure $[N_1\ [A\ N_2]_{NP_2}]_{NP_1}$ (as instantiated in the NAC_{adj} and the PGC_{adj}), the adjective is morpho-syntactically ambiguous between a determiner and a modifying adjective. To show this, the strong/weak inflection patterns of adjectives needs to be taken into account. In NPs with a strongly inflected determiner, attributive adjectives inflect according to the massively syncretistic *weak* pattern. If there is no determiner (as is the case in the alternating constructions), on the other hand, attributive adjectives inflect like determiners themselves. This is called the *strong* inflectional pattern. Thus, the adjectives in the NAC_{adj} and the PGC_{adj} have properties of adjectives as well as determiners. On the one hand, they are lexical adjectives and function as attributive modifiers. On the other hand, they are inflected like determiners, and they are the leftmost element in the NP, which is typical of determiners. This unusual double nature of adjectives in NPs without determiners leads to a plausible probabilistic interpretation of the pattern shown in Table 1. Whenever speakers classify the adjective in the kind noun phrase more as a determiner, they have to use the PGC_{adj} because, if there is a determiner, the PGC is the only option. When they classify the adjective more as an adjective, the kind NP has no determiner, and they have to use the NAC_{adj}.[8]

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

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

the NAC_{adj}. This basic attraction effect will be quantified in the corpus study Section 3.

In a probabilistic framework like Prototype Theory, the crucial question (beyond simple lemma preference effects) is, however, what controls speakers' decisions to use either variant. In the remainder of this section, I argue that the NAC and the PGC are prototypes associated with different degrees of the grammaticalisation of the measure noun and related morpho-syntactic properties, as well as register effects. The degree of similarity of a given instance to either of the two prototypes makes speakers chose the NAC_{adj} or the PGC_{adj}.

I first focus on grammaticalisation. It is often assumed that pseudo-partitives arise as a form of grammaticalised partitives (e. g., Koptjevskaja-Tamm, 2001, 536–539 for Finnish and Estonian, Koptjevskaja-Tamm, 2001, 559 for European languages in general). The grammaticalisation paths uncovered by Koptjevskaja-Tamm (2001, esp. 526–530) are relevant in the case at hand. The grammaticalisation path can start out (in some languages) with constructions involving two referential nouns (not necessarily forming a single and contiguous NP) and a *separative* meaning as in *(cut) two slices from the cake* (Koptjevskaja-Tamm, 2001, 535). The *part-of* meaning of true partitives as in *a slice of the cake* represents the first stage of a development wherein the measure noun can already lose some semantic content, when for example words like *bite* are no longer necessarily interpreted as a piece bitten out of something. The pseudo-partitive stage finally instantiates a *quantity-of* relation, potentially even leading to fully grammaticalised quantifiers such as *a lot*. In German, the PGC is clearly the older variant (Zimmer, 2015). It still has the potential to form a true partitive (if the kind noun is definite). Conversely, the NAC completely lacks this ability to form true partitives, and it thus forms the more prototypical environment for expressing measurement (as opposed to partitivity). Hence, we can expect the NAC_{adj} to be a prototypical hosting construction for more strongly grammaticalised measure nouns. For example, highly grammaticalised non-referential nouns like *Gramm* 'gram' and *Meter* 'metre' should occur proportionally more often in the NAC_{adj} than in the PGC_{adj}.

As described above, the grammaticalisation path leads from NPs denoting individuated objects standing in a *part-of* relation to a construction with a more diffuse *quantity-of* relation. Both types of relations can be numerically quantified – inasmuch as a precise number of *parts* or a numerically exact *quantity* can be specified. However, it is much more prototypical of quantities to be specified with numerical precision. Since the NAC_{adj} is more closely associated with the *quantity-of* relation, cardinals as attributes of the measure noun are expected to have a higher proportional frequency in the NAC_{adj}. For illustration, (8) shows the expected variants under this hypothesis.

(8) a. [[Drei Centiliter]$_{\text{Nom}}$ [heißer Rum]$_{\text{Nom}}$]$_{\text{Nom}}$ sind genug.
three centilitres hot rum are enough
Three centilitres of hot rum is enough.

b. [[Einige Centiliter]$_{\text{Nom}}$ [heißen Rums]$_{\text{Gen}}$]$_{\text{Nom}}$ sind genug.
some centilitres hot rum are enough
A few centilitres of hot rum is enough.

In (8a), the measure noun is modified by a cardinal *drei* 'three', and hence the NAC$_{\text{adj}}$ is preferred. In (8b), the measure noun is modified by a non-cardinal determiner *einige* 'some', and the PGC$_{\text{adj}}$ is preferred. Especially with abstract physical measure nouns (like *centilitre*), exact numerical quantification is invited. The statistical model presented in Section 3 will be specified in a way that it can detect such a preference.

Finally, it was found that the PGC$_{\text{adj}}$ is more typical of higher registers or even exclusive to written language (see Hentschel, 1993, 320–323). This is not surprising inasmuch as the genitive – an intrinsic part of the PGC – is generally underrepresented in colloquial vernacular variants of German as a result of a diachronic process wherein many (but by no means all) uses of the genitive are replaced by other cases or periphrastic constructions (Fleischer & Schallert, 2011). Under an integral view of prototypes, which incorporate effects related to larger contexts and registers, such preferences can be part of what defines the construction prototypes. In the corpus study in Section 3, register effects will therefore be modelled – even if only using two very simple proxy variables.

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

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

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

3 Corpus study

3.1 Corpus choice and sampling

For the present study, I used the German *Corpus from the Web* (COW) in its 2014 version DECOW14A (Schäfer & Bildhauer, 2012; Schäfer, 2015 and Biemann et al., 2013; Schäfer & Bildhauer, 2013 for overviews of web corpora in general and the methodology of their construction), which contains almost 21 billion tokens.[9] I chose this corpus for two main reasons.[10] 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.[11] Second, it was already mentioned that normative grammars often adopt clear positions regarding the grammaticality of either the $\mathrm{NAC}_{\mathrm{adj}}$ or the $\mathrm{PGC}_{\mathrm{adj}}$. Thus, newspaper text or any other text that conforms strongly to normative grammars might not represent the alternation phenomenon fully (and without bias) because authors and proof-readers might favour one alternative or the other. Web corpora, on the other hand, contain at least some amount of non-standard language from forums and similar sources. For these or similar reasons, COW corpora have been used in a number of peer-reviewed publications, for example Goethem &

9 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 had already been released.

10 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 irreproducibility 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).

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

Hiligsmann (2014); Goethem & Hüning (2015); Müller (2014); Schäfer (2016 aop); Schäfer & Sayatz (2014, 2016); Zimmer (2015). Therefore, DECOW can be considered the obvious choice for this study.

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

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

In step (i), I exported a list of all nouns in the DECOW14A01 sub-corpus sorted by their token frequency and manually went through it from the most frequent noun downwards, selecting the first one hundred mass nouns that occurred in the list.[12] 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 resulting list of mass nouns was used in step (ii) to bootstrap a list of measure nouns co-occurring with the mass nouns. In order to generate this list, I utilised the fact that a direct sequence of two nouns almost always instantiates the bare-noun NAC if the second noun is a mass noun. Hence, I searched for all sequences N_1N_2 where N_2 was one of the mass noun lemmas extracted in step (i). Then, the resulting 100 lists of noun-noun combinations were each sorted by frequency in descending order and sieved manually to remove erroneous hits. From each of the 100 lists, I also removed noun-noun combinations that had a frequency below 2, except if the individual list would have otherwise been shorter than 20 noun-noun combinations. The result was a list of the most frequent 2,365 individual combinations of a measure noun and a mass noun.

12 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 ($\mathrm{PGC}_{\mathrm{adj}}$ and $\mathrm{NAC}_{\mathrm{adj}}$) individually in each of the first ten slices of DECOW (roughly 10 billion tokens). In order to reduce the sample size for the manual annotation process, the final concordance was sampled from the results of these 2,365 queries. Since the mass nouns in the sample were distributed according to the usual power law, I used all hits for 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 $\mathrm{NAC}_{\mathrm{adj}}$ cannot be distinguished from the $\mathrm{PGC}_{\mathrm{adj}}$. Given the careful bootstrapping and sampling procedure described in this section, we can be highly sure that it contains all relevant and reasonably frequent noun–noun combinations in the target constructions.[13]

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

3.2 Variables and annotation

The full set of manually annotated variables for the main sample is given in Table 2, and I briefly discuss it now.[14] 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 $\mathrm{PGC}_{\mathrm{det}}$ and the $\mathrm{NAC}_{\mathrm{bare}}$. They were calculated from the auxiliary samples described at the end of Section 3.1 as a log-transformed quotient. The higher the value,

13 In a similar fashion, the 100 most frequent measure nouns occurring with plural kind nouns were bootstrapped and queried, resulting in a sample of 871 sentences. As stated in Section 2, the $\mathrm{NAC}_{\mathrm{adj}}$ 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 $\mathrm{NAC}_{\mathrm{adj}}$ and 794 times $\mathrm{PGC}_{\mathrm{adj}}$). The sample is distributed in the data package accompanying this paper, however.

14 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).[15] 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 variants. To capture this, class information was annotated for measure lemmas. The classification was inspired by the list in Koptjevskaja-Tamm (2001, 530), but due to the low frequencies of many of the potential classes, a very coarse classification was finally used. With typical examples and their frequencies in the final sample, the classes are: *Physical* (abstract precisely measurable units such as *Liter* 'litre', *Meter* 'metre', *Gramm* 'gram'; $f = 1,968$), *Container* (*Eimer* 'bucket'; $f = 740$), *Amount* (*Menge* 'amount'; $f = 1,364$), *Portion* (natural portions like *Happen* 'bite' or *Krümel*

15 It could be argued that some more advanced measure of attraction strength should be used, as is done in Collostructional (or Collexeme) Analysis (Gries & Stefanowitsch, 2004), see also Gries (2015b). Three main points speak against such an approach in the present case (but see Levshina, 2016, 246–249 for a different approach). First, the attraction values will be used as regressors in a hierarchical logistic regression, and the values resulting from collostructional approaches, 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 $-\infty$. Second, their main use in collostructional analyses is to *sort* a list of collexemes and interpret their order. Their concrete numerical value might not have a solid cognitive interpretation in the given context. Third, I tried using collexeme strength as a regressor, and the results were unsatisfactory.

'crumb'; $f = 713$). The few 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 $\mathrm{NAC}_{\mathrm{adj}}$ as hypothesised in Section 2.2.

To capture the influence of register or style mentioned in Section 2.2, two proxy variables were used. At the document level, the DECOW corpus has an annotation for *Badness*. As described in Schäfer et al. (2013), *Badness* measures how well the distribution of highly frequent short words in the document matches a pre-generated language model for German. Documents with higher Badness usually contain more incoherent language, shorter sentences, etc. If the $\mathrm{PGC}_{\mathrm{adj}}$ actually favours higher registers and styles, 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. While more genitives are also indicative of higher registers, the use of this variable as a regressor in the present study might be considered problematic. Since the $\mathrm{PGC}_{\mathrm{adj}}$ contains a genitive itself, the regressor variable *Genitive* and the document-level variable *Genitives* are not fully independent. However, since instances of the $\mathrm{PGC}_{\mathrm{adj}}$ 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 $\mathrm{PGC}_{\mathrm{adj}}$ more than the corresponding nominative and accusative MNPs (Hentschel, 1993; Zimmer, 2015). As an example, *mit einem Stück frischen Brots* 'with a piece of fresh bread' ($\mathrm{PGC}_{\mathrm{adj}}$) would be preferred over *mit einem Stück frischem Brot* ($\mathrm{NAC}_{\mathrm{adj}}$). As with all the examples, native speakers of German will most likely notice that differences are subtle. To control for this effect, the case of the measure noun was manually annotated (variable *Measurecase*).

3.3 On statistical analysis

In this section, I justify the choice of statistical models which I use 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/

GL(M)M. Over the past few years, however, modified or alternative methods have been proposed. From among these methods, 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.

As for the first point, this is not quite the place to discuss it fully. The basic distinction is a philosophical one and related to the concepts of *direct* and *inverse probability* (e. g., Senn, 2011). Frequentists assume that models and parameters are fixed, for example a model specifying that a coin is fair. We can then calculate for observed data (for example a measurement of 3 heads in an experiment with 10 tosses) how often such a result or a more extreme result would occur if the model were true and we repeated the experiment arbitrarily often. This is essentially the frequentist notion of direct probability, i. e., long-run frequencies under replication. Standard tests in the Fisher and Neyman-Pearson traditions as well as Neyman confidence intervals are based on this concept of probability. Bayesian approaches (in the now common interpretation), on the other hand, condition on the particular data and quantify inductively the probability of model parameters given the data. The parameters are thus not fixed, and the probability is usually equated with researchers' posterior beliefs about model parameters. There is actually a debate among Bayesians about the proper interpretation of Bayesian methods, and whether a notion of hypothesis testing is compatible (or even already contained) in the Bayesian approach. In Gelman & Shalizi (2013, 10), the authors – prominent Bayesians themselves – acknowledge that a theory of statistical testing is a desideratum, and they state about the standard inductive interpretation of Bayesianism that "most of this received view of Bayesian inference is wrong", and they develop a Bayesian notion of p values (see also Mayo, 2013, for a frequentist reply; also Senn, 2011 on different strands of Bayesianism and their stance on inductive vs. deductive reasoning, and Mayo, 2011, for a critical reply to Senn, 2011). Clearly, in such quarrels between and among camps of philosophers of science and statisticians, it is difficult for mere practitioners to take sides.

These quarrels relate to the second point, however. Divjak (2016, 301–302) speaks favourably of Bayesian methods because the Bayesian concept of probability is more adequate for cognitive modeling compared to the frequentist one.

Her argument is part of a larger body of literature asking for cognitively plausible modeling techniques, for example Naive Discriminative Learning (NDL; Baayen, 2011; Baayen et al., 2013; Milin et al., 2016; Theijssen et al., 2013). On p. 303 of Divjak (2016), the author goes on to explicitly mention NDL as well. Yet, neither frequentist nor Bayesian methods were conceived as cognitive models, but as systems of inference for scientists (see above, and see also Divjak, 2016, 302). The fundamental question that lurks behind such arguments is how we interpret our statistical models (estimated on corpus data). Are they inductive models of cognitive representations that also human learners would infer from being exposed to the corpus data?[16] Or are they tests of theories that are pre-specified and merely tested for predictive accuracy on linguistic output data contained in corpora? In the former case, we adopt a strong *corpus as input* hypothesis and should definitely resort to methods like NDL. However, this would most likely require us to toss most previous work done in (cognitive) linguistics into the bin and to abandon all high-level generalisations that most existing studies have been based upon. In fact, arguments to this extreme effect and against using high-level generalisation have been made, for example in Baayen et al. (2016), Divjak (2016, 299–300), Ramscar & Port (2016), Theijssen et al. (2013). In the latter and less extreme case, the cognitive commitment does, however, not necessarily extend to the statistical methods used. These methods then do not need to be any more cognitively plausible than an ANOVA used to analyse the results from an experiment. I view my own work (see Sections 2–4) in the tradition of testing theories (which embrace high-level generalisations), and I agree to provisionally use Likelihood methods (see below). I remain fully agnostic with respect to the question, which approach is *right*. The best strategy for cognitive linguistics as a field might be to cultivate many methods while making sure that each method is applied carefully and competently.

The third point, then, is the most practically relevant in the context of this paper. In her otherwise excellent study, Levshina (2016, 251–252) argues for Bayesian estimation in mixed regression settings. First, she claims that "while frequentist statistics only allows one to test whether the null hypothesis can be rejected, Bayesian statistics enables one both to test the null hypothesis and to estimate the probability of specific parameter values given the data." This does not do justice to frequentist methods in that mere rejection of the null hypothesis is characteristic only of Fisher's approach. In the Neyman-Pearson approach,

16 In which case we are doing "data science in language research" in the words of Milin et al., 2016. I see this as standing in contradiction to the view advocated in Dąbrowska (2016) as cited above.

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.[17] Now, Maximum Likelihood Estimation (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 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). Going back to the argument that Bayesian methods are allegedly more cognitively plausible than 'frequentist' methods, the claim might still be true. However, most similar studies (dating back to Bresnan et al., 2007) use an estimator (MLE) that is not frequentist in a narrow sense, and which is as cognitively plausible as Bayesian estimators because it most likely produces the same results in these studies. If for Levshina's study, results from Bayesian and MLE methods did, in fact, *not* converge, it would have been an ideal occasion to demonstrate the 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

17 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.

separation, etc. (see Freedman, 1999, Gelman & Hill, 2006, 345–348). I want to reiterate 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. That said, I want to voice the concern that probably, many practitioners are 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 Section 3.4, I compare my own hierarchical models estimated with Bayesian and MLE estimators 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* (Bates et al., 2015) for Maximum Likelihood Estimation, and *rstanarm* (Gabry & Goodrich, 2016) for 'Bayesian' Markov-Chain Monte Carlo 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 $1,1344$ cases of $\mathrm{PGC}_{\mathrm{adj}}$ and $3,929$ cases of $\mathrm{NAC}_{\mathrm{adj}}$. The results of the estimation are shown in Table 3 and in Figure 1. The regressors with the measure lemma as their unit of reference have no within-measure lemma variance, and the *glmer* function automatically estimates them as *group level predictors* (or *second-level effects*), cf. Gelman & Hill (2006, 265–269,302–304). The same goes for those listed with the kind lemma as their unit of reference. Given the coding of the response variable, coefficients leaning to the positive side can be interpreted as favouring the $\mathrm{PGC}_{\mathrm{adj}}$.

Standard diagnostics for 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 $\mathrm{GVIF}^{1/2\mathrm{df}}$ was higher than 1.6. Nakagawa & Schielzeth's pseudo-coefficient of determination is $R^2_m = 0.409$ and $R^2_c = 0.495$ (see Gries, 2015a for a 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 $\sigma_{\mathrm{Measurelemma}} = 0.448$ and $\sigma_{\mathrm{Kindlemma}} = 0.604$.

The coefficient estimates are specified in Table 3 for each regressor (or regressor level) in the columns labelled *Coefficient*. For a robust quantification

Level	Regressor	p_{PB}	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 (*Cardinal=Yes*, *Measurecase=Nom*, *Kindgender=Masc*, *Measureclass=Physical*; 0 for all numeric z-transformed regressors) is -3.548 (MLE) and -3.700 (MCMC)

of the precision of the estimation, I ran a parametric bootstrap (using the *confint.merMod* function from *lme4*) with 1,000 replications, and using the percentile method for the calculation of the intervals. The resulting 95% bootstrap confidence intervals are reported in Table 3 in the columns labelled *CI low* and *CI high* (= upper and lower 2.5th percentiles). The column *CI contains 0* shows an asterisk for those intervals that do *not* include 0. Furthermore, for each regressor, a p value was obtained by dropping the regressor from the full model, re-estimating the nested model and comparing it to the full model. Instead of inexact Wald approximations or Likelihood Ratio Tests, I used a drop-in bootstrap replacement for the Likelihood Ratio Test from the function *PBmodcomp* from the *pbkrtest* package (Halekoh & Højsgaard, 2014). I call the corresponding value p_{PB}, and it is given in the appropriately labelled columns in Table 3.

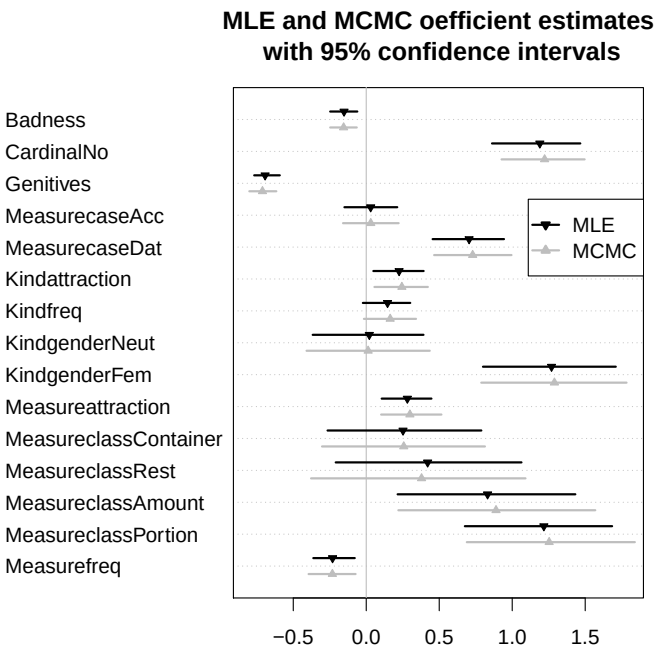

Fig. 1: Coefficients (MLE and MCMC) with 95% confidence intervals (for details see text); the intercept (*Cardinal=Yes*, *Measurecase=Nom*, *Kindgender=Masc*, *Measureclass=Physical*; 0 for all numeric z-transformed regressors) is -4.328 (MLE) and -4.441 (MCMC)

Only *Kindfreq* ($p_{\text{PB}} = 0.095$) can be seen as slightly too high to be convincing (non-significant).

Finally, I show that Bayesian methods in the form of Markov-Chain Monte Carlo sampling (MCMC) do 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.[18] The algorithm converged, and for all coefficients, the $\widehat{R}$ diagnostic was exactly 1. The resulting coefficients and intervals as well as an * are also given in Table 3 in the columns labelled *MCMC*. Both methods lead to exactly the same results (minus negligible numerical differences) as expected given the modest complexity of the model structure and the large sample size (see Section 3.3). The signs and magnitudes of the coefficients are identical, and confidence intervals have the same width and symmetry properties. Figure 1 illustrates this by also showing both estimates. Since both estimators converge, I only interpret the MLE model in the next section.

3.5 Interpretation

The results reported in Section 3.4 generally confirm the hypotheses from Section 2.2. First, the prototypicality effect related to the non-alternating PGC_{det} and NAC_{bare} can be shown (see the effect plots in Figure 2).[19] The effect is mostly as expected: if a lemma appears relatively more often in the PGC_{det} (compared to its frequency in the NAC_{bare}), the more often the PGC_{adj} is cho-

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

19 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 grey areas (continuous variables) are asymptotic 95% confidence intervals calculated from *glmer*. They are not bootstrapped.

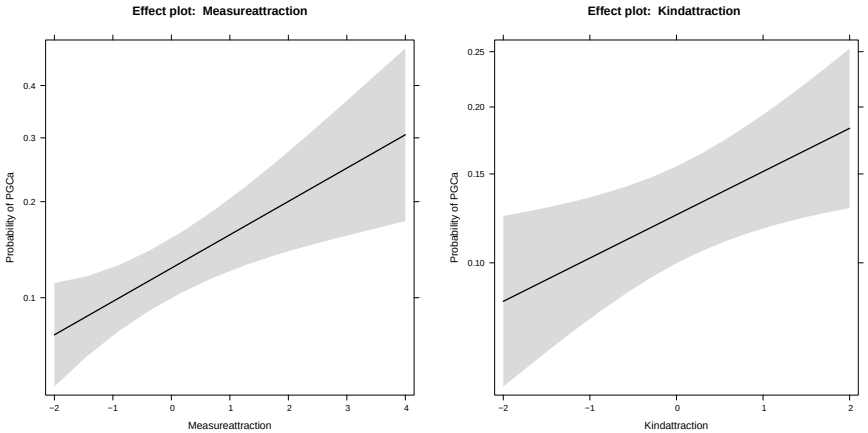

Fig. 2: Effect plots for the regressors *Measureattraction* and *Kindattraction*; y axes are not aligned

sen 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 $\mathrm{NAC}_{\mathrm{adj}}$ ($\beta_{\mathrm{Measurefreq}} = -0.231$, $p_{\mathrm{PB}} = 0.005$) 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.

In Section 2.2, it was also hypothesised that classes of measure nouns with a higher degree of grammaticalisation should favour the $\mathrm{PGC}_{\mathrm{adj}}$. The *Measure-class* 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).

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

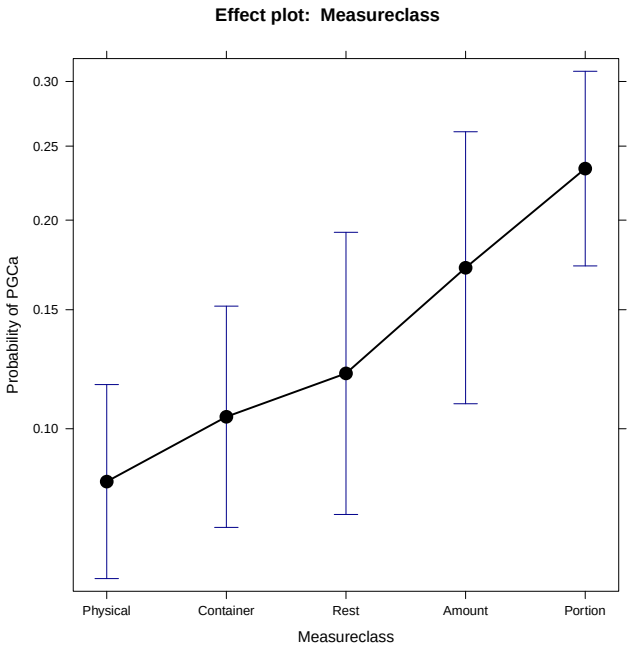

Fig. 3: Effect plot for the regressor *Measureclass*

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

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

The influence of *Measurecase* ($p_{\mathrm{PB}} = 0.001$) is as predicted in previous analyses (see Section 2.2). A measure noun in the dative favours the $\mathrm{PGC}_{\mathrm{adj}}$ with $\beta_{\mathrm{MeasurecaseDat}} = 0.705$ (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.

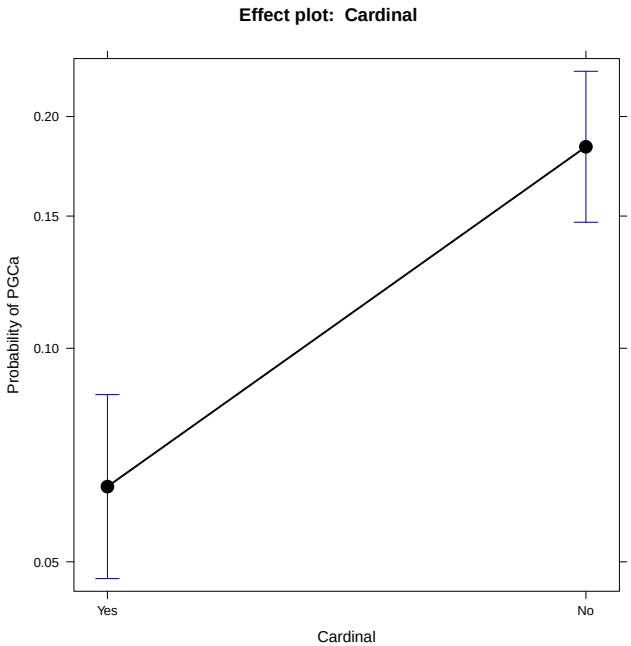

Fig. 4: Effect plot for the regressor *Cardinal*

4 Correlating corpus findings with 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 variants in these sentences is 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 variants, which is intended to produce more subtle results compared to a

two-alternative forced-choice task (Rosenbach, 2013), where participants have to choose one of two variants. 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), it was shown that results from forced-choice and split-100 experiments mostly converge, which is expected in the long run if a probability is turned into a binary decision. 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 the task of distributing a probability mass across two variants. Participants had to chose between two sentences differing only in that one contained the NAC_{adj} and the other contained the PGC_{adj}. The analysis compares the probabilities assigned to stimuli by the corpus-derived model with the frequency with which participants chose the variants for the same stimuli.

There were 24 participants (native speakers of German without reading or writing disabilities) aged 19 to 30 living permanently in [name of city anonymised], who were recruited from introductory linguistics courses at [name of university anonymised]. Although the experiment was conducted in the last four weeks of their first semester, participants had no deeper explicit knowledge of linguistics, grammar, or experimental methods. None of them had ever participated in a forced-choice experiment before. Participation was voluntary, but participants received credit in partial fulfillment of course requirements.

As stimuli, attested MNPs from the corpus study were used, but the sentences were radically simplified to avoid influences from contextual nuisance factors as much as possible. The approach is also justified because according to the theoretical assessment in Section 2.2, the choice of variants depends mostly on a very local constructional context. I sampled 16 MNPs from the corpus, and it was made sure that the simplifications and normalisations did not affect any of the regressors used in the corpus study. In the simplified sentences, the case, number, etc. of the MNP remained the same as in the attested sentence, as well as the choice of lexical material within the MNP. Eight sentences contained masculine or neuter kind nouns, the other eight contained feminine kind nouns. Furthermore, in each of the masculine/neuter and feminine groups, four sentences originally containing the 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.[20] High and low

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

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

probability were defined as the top and bottom 20% of all probabilities assigned by the GLMM. Lemmas and feature combinations were made unique within each group whenever possible. The design is summarised in Table 4.

The pairs of stimuli were the sentence containing the preferred construction (according to the corpus GLMM) and a modified version containing the dispreferred construction. They were presented next to each other, and a 20 second time limit for each choice was set.[21] 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 alternation phenomena from German morphosyntax were used. Thus, participants saw 39 pairs of sentences and 78 sentences in total. They were instructed to select from each pair of sentences the one that seemed more natural to them in the sense that they would use it rather than the other one. The experiment was conducted using *PsychoPy* (Peirce, 2007).

Then, a multilevel logistic regression was specified with the probability of the $\mathrm{PGC_{adj}}$ predicted for each sentence by the corpus-based GLMM as the only fixed effect *Modelprediction*.[22] 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.[23] Coefficients were estimated with Maximum Likelihood Estimation (*lmer* function from *lme4*). The number of observations was $n = 384$.

A certain amount of the variance can be accounted for by idiosyncrasies of single sentences ($\sigma_{\mathrm{Sentence}} = 1.785$, $\sigma^{\mathrm{Modelprediction}}_{\mathrm{Sentence}} = 5.996$, 16 levels).[24] Also,

21 No participant ever exceeded the time limit.

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

23 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.

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

among participants, there are clearly different preferences ($\sigma_{\text{Participant}} = 0.781$, $\sigma^{\text{Modelprediction}}_{\text{Participant}} = 0.484$, 24 levels). On the extreme sides, one participant chose the 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} variant 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.

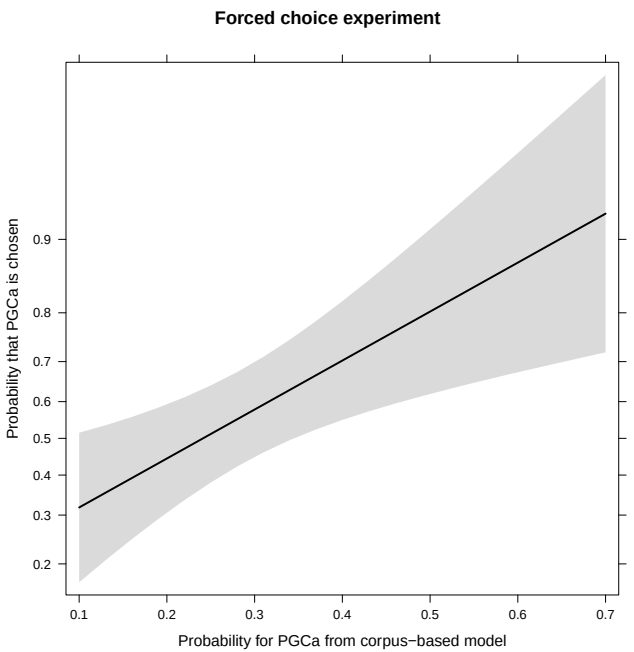

Fig. 5: Effect plot for the multilevel logistic regression in the forced-choice experiment: predictability of participants' choices using the probabilities derived from the corpus-based GLMM

4.2 Experiment 2: self-paced reading

The second experiment tests preferences more implicitly. It is expected that reading less prototypical variants (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 variant predicted by the corpus model and the 16 modified sentences with the variant that the corpus model did not predict.[25] 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 and so was the order within each block. The sentences from each pair of variants were kept as far apart as possible. The fillers also came in pairs such that the second block exclusively contained sentences to which participants had been exposed in the first block in slightly modified form. In total, each participant read 98 sentences. After each sentence, participants had to answer simple (non-metalinguistic) yes-no questions about the previous sentence as distractors. The distractor questions were different between the first and the second blocks. There were 38 (different) 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, such as the bracketed words in the example *zwei Gläser* [*sprudeln-*

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

Regressor	Coefficient	CI low	CI high	0 not in CI
ConstructionPGCa	0.054	0.012	0.095	*
Modelprediction	-0.006	-0.113	0.110	
Position	-0.005	-0.005	0.004	
ConstructionPGCa:Modelprediction	-0.125	-0.234	-0.023	*

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

des Wasser] 'two glasses of sparkling water'. Outliers farther than 2 inter-quartile 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 PGC_{adj}. As a consequence, the higher the GLMM prediction is, the more prototypical the sentence is for containing the PGC_{adj}. Therefore, it is expected that reading times are higher when model *Modelprediction* is higher and the sentence contains the NAC_{adj}. However, when the sentence contains the PGC_{adj}, reading times should be lower when *Modelprediction* 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).[26]

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 $R^2_c = 0.346$.

26 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.

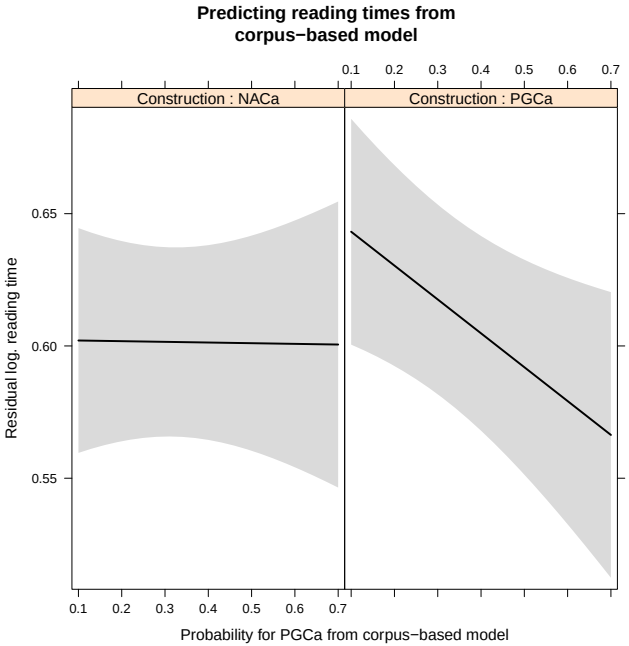

Fig. 6: Effect plot for the LMM in the self-paced reading experiment: modeling participants' residualised log reading times on the probabilities given by the corpus-based GLMM

The overall model quality is high, and the effect plot for the main effect 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 as the construction with the lower frequency might be developing towards a more sharply defined prototype.[27] Conversely, the NAC_{adj} (like the NAC in

27 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).

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.

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 brand of cognitive linguistics apart from *both* Aristotelian frameworks of the generative flavour *and* introspective, non-empirical, and anti-quantitative versions of cognitive linguistics (see Dąbrowska, 2016 for an impressive 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 variants in line with the corpus-derived model. In a more implicit self-paced reading experiment, it was shown that the much rarer variant brings about a reading time penalty except in cases for which the corpus model predicts very high probability for this variant.

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.

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.

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

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

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

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

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

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.

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

Eisenberg, Peter. 2013. *Grundriss der deutschen Grammatik: Der Satz*. Stuttgart: Metzler 4th edn.

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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

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

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

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

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

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

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

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

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

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

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

Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt. 2010. The German reference corpus DeReKo: A primordial sample for linguistic research. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC '10)*, 1848–1854. Valletta, Malta: European Language Resources Association (ELRA).

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

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

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

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

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

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

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

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

Mayo, Deborah G. 2011. How can we cultivate Senn's ability? *Rationality, Markets, and Morals* 3. 14–18.

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

Medin, Douglas L. & Marguerite M. Schaffer. 1978. Context theory of classification learning. *Psychological Review* 85(3). 207–238.

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

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

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

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

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

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

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

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

Rosch, Eleanor. 1978. Principles of categorization. In Eleanor Rosch & Barbara B. Lloyd (eds.), *Cognition and categorization*, 27–48. Hillsdale: Lawrence Erlbaum Associates.

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.

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

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

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

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

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

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

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

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

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

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

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

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

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.

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

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

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.

Verhoeven, Elisabeth & Anne Temme. 2017. Word order acceptability and word order choice. Submitted.

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

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

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

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