Roland Schäfer*

Abstractions and exemplars: the measure noun phrase alternation in German

Abstract: In this paper, an alternation in German measure noun phrases is examined under a varying-abstraction perspective. In a specific measure NP construction, the embedded kind-denoting noun either agrees in case with the measure noun (*eine Tasse guter Kaffee* 'a cup of good coffee') or it stands in the genitive (*eine Tasse guten Kaffees*). Each of the two alternants is syntactically similar to a non-alternating construction. I propose a prototype model which assigns a common prototypical meaning to each of the alternants and its corresponding non-alternating construction. Based on this, I argue that lexical, morpho-syntactic, and stylistic features help to predict the choice of the alternant. A large corpus study is presented which supports this analysis. However, in addition to the prototype effects, an exemplar effect is also shown to influence the choice, namely the relative frequencies with which lemmas occur in the non-alternating constructions. I argue that allowing both prototype and exemplar effects is more adequate than following radical prototype or exemplar approaches. It is also verified in two experiments that the corpus-derived model corresponds to the behaviour of native speakers. The weak effect size of the experimental validation is discussed in the context of corpus-based cognitive linguistics and the validation of corpus-derived models.

Keywords: prototypes vs. exemplars, corpus methods and experimental validation, alternations, multilevel modeling, pseudo-partitives, German

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

1 Prototypes, exemplars, and corpora in cognitive linguistics

1.1 Alternations

This paper deals with a morpho-syntactic alternation between two measure noun constructions in German. By *alternation* I refer to a situation where two or more forms or constructions are available with no clear difference in acceptability, function, or meaning. The study of lexical and constructional alternations has a long history in cognitive corpus linguistics (for example, Bresnan et al., 2007; Bresnan & Hay, 2008; Bresnan & Ford, 2010; Divjak & Arppe, 2013; Gries, 2015b; Nesset & Janda, 2010). This research is based on the assumption that language is a probabilistic phenomenon where alternants are chosen neither deterministically nor fully at random (Bresnan, 2007). Multifactorial models are constructed which incorporate influencing factors from diverse levels, including lexical and contextual factors. The estimation of the model coefficients quantifies the influence that the factors have on the probability that either alternant is chosen. There are two fundamental issues to consider with respect to this tradition. First, there is the question of whether corpus data provide any insight into cognitive representations at all. This question can and should be answered by testing how well corpus-derived models converge with or diverge from experimental findings (Secion 1.2). Second, the appropriate modelling of such results in cognitive linguistics – i. e., the assumed constructs – is a key issue (Section 1.3).

1.2 Usage data in cognitive linguistics

For some time, there has been an interest in correlating probabilistic generalisations extracted from corpus data and results from experimental work (for example, Arppe & Järvikivi, 2007; Bresnan et al., 2007; Bresnan & Ford, 2010; Divjak & Gries, 2008; Divjak et al., 2016a; Ford & Bresnan, 2013). This is often called a *validation* of the corpus-derived findings, but Divjak (2016, 303) 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". See also Newman (2011) for a positive view of corpora as a source of data in cognitive linguistics in their own right. However, as Dąbrowska (2016, 486–487) argues, this does not mean that we can in some way "deduce mental representations from patterns of use", i. e.,

from corpus data. It would indeed be highly surprising if this were possible, and the same holds for experimental methods. Nobody assumes that we can inductively infer mental representations from experiments, which even allow for direct access to the cognitive agent and offer much better ways of controlling experimental conditions and nuisance variables. Rather, predictions are derived from existing theories in order to *test* the theories. While this approach is applicable to corpus data, I discuss some relevant differences.

As mentioned above, the central question is whether usage data as found in corpora are truly predictive of speakers' and writers' cognitive representation of language and/or of their linguistic behaviour. This is where the experimental validation (or *corroboration*) comes into play. A review of the state of the art was provided by Newman & Sorenson Duncan (2015), who enumerate a number of studies showing how corpus data and experimental data converge (such as Bresnan et al., 2007; Durrant & Doherty, 2010; Gries & Wulff, 2005; Gries et al., 2005) and a number of studies where the two types of data led to diverging or only partially converging results (such as Arppe & Järvikivi, 2007; Dąbrowska, 2014; Mollin, 2009). When researchers do not achieve convergence, they often try to explain this by differentiating between the actual cognitive construct and whatever the pooled usage data as found in corpora represent. For example, Dąbrowska (2014, 411) lists a number of possible reasons to explain why subjects in her experiment diverged in their word association preferences from collocation measures extracted from corpora. Alternatively, researchers argue for a more adequate statistical analysis to increase the fit between corpus data and experimental data. See, for example Divjak et al. (2016a), who show that generalised additive models (GAMs) are better suited than generalised linear models (GLMs) for correlating reading times and corpus data. Much more optimistically, Stefanowitsch & Flach (2016) proposed a straightforward positivist perspective of corpora as representing the input of an average adult speaker, thus licensing inferences from corpus data to cognitive representations under a "corpus-as-input" view. They state that "in this wider context, large, register-mixed corpora such as the British National Corpus [...] may not be perfect models of the linguistic experience of adult speakers, but they are reasonably close to the input of an idealized average member of the relevant speech community" (Stefanowitsch & Flach, 2016, 104). In essence, this entails that any generalisation extracted from the BNC can be assumed to have some kind of mental reality, which is at least doubtful.

No general consensus has emerged yet, which is not surprising given the number of cognitive constructs assumed at diverse levels, the problems of corpus composition, the operationalisations involved in experiments, and the choice of statistical tools. While far from providing definitive solutions, my discussion in

Section 5 will provide possible explanations for the quality of the fit between the corpus data and the experimental data reported in Sections 3 and 4, much in the spirit of Dąbrowska (2014).

1.3 Prototypes, exemplars, corpora, and controlled experiments

I now turn to cognitive representations as constructs in alternation research. The typical approach in alternation research is to annotate a large number of corpus sentences with linguistic features and to model the probability of the variants being chosen given these features. The idea is that a variant is chosen when the influencing features have certain typical values. In other words, a variant of prototype theory with features (Rosch, 1978) is the appropriate model, also because the features used to feed the model are usually abstract high-level linguistic features. Some researchers such as Gries (2003), Nesset & Janda (2010), or Schäfer (2016) indeed commit to prototype theory in alternation modelling. Under prototype theory, category membership is defined by similarity to an ideal exemplar and its characterising features (see Taylor, 2008 for an overview).

While prototype theory is well suited for modelling constructional choices, it is just one of at least two major similarity-based theories of classification, the prominent alternative being exemplar theory (Medin & Schaffer, 1978; Hintzman, 1986). Prototype theory and exemplar theory model essentially the same effects when only output data are considered. The theories differ significantly in whether they assume higher-level abstractions in the form of single maximally prototypical exemplars (prototype theory) or categories emerging through the storage of many exemplars and similarity-based classification over those sets of exemplars (exemplar theory). Barsalou (1990) already showed that prototype and exemplar theory model the same types of surface effects and are informationally equivalent. Consequently, experiments which favour one theory over the other use procedural behaviour of subjects in experiments, for example the speed of category retrieval, as opposed to mere output data.[1] Since corpus data only show artefacts of production events and we have no experimental access to the speaker's or writer's performance and their actual similarity judgements, one should be sceptical whether corpus analysis alone could ever help to decide which theory of mental representation is more suitable (see also Gries, 2003,

1 See Storms et al. (2000) for a comparison of the theories in different experimental settings.

22 and the Dąbrowska, 2016, 486–487 quote above). However, as I will show, some effects are naturally analysed as prototype effects while others are almost necessarily exemplar effects.

In cognitive science, it is mostly accepted that exemplar theories have greater explanatory power (Vanpaemel, 2016, 184), and that abstraction is only needed marginally, if at all. Still, various attempts have been made to settle the dispute between abstraction-based models (models with rules or prototypes) and exemplar models. Vanpaemel & Storms (2008) and Lee & Vanpaemel (2008) proposed the *varying abstraction model* (VAM) which "attempt[s] to balance economy and informativeness" (Lee & Vanpaemel, 2008, 745), treating models with full abstraction (radical prototype theory) and no abstraction at all (radical exemplar theory) as special cases of a flexible approach which allows both types of mechanisms. The mixture model of categorisation (MMC) by Rosseel (2002) is a model with abstraction in the form of hierarchical clusters of exemplars. These clusters of objects are characterised by a probability distribution over their features, and categorising new objects is a process of estimating the probability of this object belonging to one of the clusters. Griffiths et al. (2009) go further and present a computational model which is able to choose the appropriate complexity of representation for a given category. However, despite these and more attempts to reconcile or unite the two approaches Vanpaemel (2016, 183–184) describes the state of affairs between adherents of neo-prototype theory (such as Minda & Smith, 2001, 2002) and exemplar theory as a stalemate.

In cognitive linguistics, similarity-based categorisation is often seen as the central and conceptually sufficient approach to linguistic categorisation. See Taylor (2003) for a comprehensive treatment in terms of prototype theory and Taylor (2012) for a detailed discussion of an exemplar-based approach. Divjak & Arppe (2013) is a rare example of a paper in cognitive linguistics where a synthesis between prototype and exemplar models is proposed. Their corpus-based approach shows "one way of systematically analyzing usage data as contained in corpora to yield a scheme, compatible with usage-based theories of language, by which the assumptions of both the prototype and exemplar theories can be operationalized" (Divjak & Arppe, 2013, 267). Their approach to implementing a varying abstraction model (Divjak & Arppe, 2013, 254–260) is based on hierarchical clustering of annotated properties of sentences. They cluster sentences containing Russian verbs of trying. Then, they single out the one sentence from each cluster which scores the highest probability for any of the six *try* verbs according to a polytomous regression model estimated on the same data. The clusters are interpreted as intermediate-level exemplar-derived abstractions of typical contexts for these high-probability verbs (typically more than one cluster for each verb; Divjak & Arppe, 2013, 255–256). The crucial difference between such data-driven corpus-

based analyses and experiments in cognitive science (Divjak & Arppe, 2013 use Verbeemen et al., 2007 as their reference) is that cognitive research is based on experiments where subjects produce actual category assignments or similarity judgements, and in corpus studies, the categories and category membership are determined purely from existing data. The experimental approach with reduced and/or artificial stimuli makes it much easier to examine very specific effects in the behaviour of the subjects. While I am convinced that the results presented in Divjak & Arppe (2013) are valid and important (especially given the previous and subsequent research the authors have conducted on the data, including experimental work presented in Divjak et al., 2016a), any data set can be clustered to yield a certain number of clusters. Thus, the study does not ensure that the clusters emerging from the data correspond to any speaker's cognitive representation.

On the other hand, the trade-off one has to accept when doing experiments with highly simplified stimuli is lower *external validity* (i. e., decreased generalisability) and higher dependence on proper operationalisations of constructs (*construct validity*).[2] Tasks in cognitive science have been criticised exactly for their lack of external validity, for example by Murphy (2003). From a linguistic perspective, it seems remarkable that Voorspoels et al. (2011, 1013) consider their experimental task – the assignment of typicality scores to nouns from the domains of *animals* and *artefacts* to categories like *bird*, *fish*, *clothing*, or *tools* – a study of "superordinate natural language categories, whereas most evidence supporting exemplar representations has been found in artificial categories of a more subordinate level". Corpus linguists interested in probabilistic alternation modelling investigate significantly more complex high-level categories and use large and complex feature sets.[3] It is thus an advantage of much linguistic work on categorisation that it deals with complex and realistically produced data because this greatly improves the external validity of studies, although by potentially sacrificing some construct validity. Thus, an ideal contribution to the research on category abstraction by cognitive corpus linguists is to provide analyses which have great external validity and complexity while carefully making sure that these finding correlate with actions and reactions under more controlled experimental conditions, thus increasing the construct validity.

This paper contributes to solving the questions raised in this section in many ways. After a thorough description of the alternation under examination,

2 An accessible overview of the different types of validity can be found in Chapter 1 of Maxwell & Delaney (2004).

3 However, approaches have emerged which model linguistic phenomena without reference to high-level linguistic features altogether (Baayen et al., 2016; Ramscar & Port, 2016).

I discuss potential influencing factors comprising high-level abstract semantic generalisations as well as exemplar-similarity and item-specific effects in Section 2. I will argue that there are lemma-specific and exemplar effects but also generalisations at the level of the construction as well as generalisations overlaying the lemma-specific effects, leading to a complex hierarchical structure. In Section 3, I present a corpus study and report a true multilevel generalised linear model with the appropriate hierarchical structure for the hypothesised effects. In Section 4, I test the predictions of the corpus-based model in two experimental paradigms (forced-choice and self-paced reading), showing that they indeed converge, albeit with low effect strength. In Section 5, I interpret the findings in the light of the issues of cognitive representations and corpus data and the convergence of usage data and experimental data.

2 Modelling the measure noun alternation

In this section, I introduce and illustrate the relevant alternating constructions in Section 2.1. Then, I develop a model with prototype effects as well as exemplar effects based in existing research and my own theorising in Section 2.2.

2.1 Alternating and non-alternating measure NP constructions

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

2.1.1 The measure noun phrase alternation

Three different syntactic configurations within MNPs need to be distinguished. The case alternation only occurs when the kind noun is modified by an attributive adjective and there is no determiner, as in (1). Superficially, the sentences are functionally and semantically equivalent either with the kind noun in the genitive (1a) or in the same case as the measure noun – an accusative in the case of (1b).[4] Notice that the genitive is virtually always used when the kind noun is a plural count noun, and the corresponding cases are consequently not included in the present study.

4 Some descriptive and normative grammars take stronger positions with regard to the acceptability of the two options. See Hentschel (1993) and Zimmer (2015) for analyses of the sometimes absurd stances taken in grammars of German.

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

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

This specific configuration has to be seen in the context of two other configurations, to which I turn now. First, if the kind noun forms an NP with a determiner, the construction resembles (and is usually called) a *pseudo-partitive* (on partitives and pseudo-partitives see, e. g., Barker, 1998; Selkirk, 1977; Stickney, 2007; Vos, 1999; for a recent application of the terminology to German, see Gerstenberger, 2015).[5] Here, the kind noun is in the genitive, and I refer to the construction in (2) as the *Pseudo-partitive Genitive Construction* (PGC).

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

Second, if the kind noun is bare – i. e., if it comes neither with a determiner nor a modifying adjective – it is uninflected as in (3a), and the genitive as seen in the PGC is not acceptable, see (3b).[6]

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

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

This construction is usually classified as a *Narrow Apposition Construction* (Löbel, 1986), henceforth NAC. Notice that the unavailability of the genitive

5 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).

6 It is difficult to determine whether the bare kind noun in (3a) bears no case at all, a generic case, or agrees in case with the measure noun. When there is an adjective as in (1b), the embedded kind NP clearly agrees in case, but due to the overall absence of markers of case in the singular, bare nouns mostly show no indication of their case. Descriptive grammars seem to favour an analysis in terms of caselessness (for example, Zifonun et al., 1997, 1981). Since it is not relevant to the argument presented here, the question is left open.

kind NP is:	bare noun NP [...N_{meas} [N_{kind}]]	NP with adjective [...N_{meas} [AP N_{kind}]]	NP with determiner [...N_{meas} [D N_{kind}]]
narrow apposition	NAC_{bare} (3a) *Glas Wein*	NAC_{adj} (1b) *Glas guten Wein*	—
pseudo-partitive genitive	—	PGC_{adj} (1a) *Glas guten Weins*	PGC_{det} (2) *Glas dieses Weins*

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

on the kind noun follows independently from a constraint that genitive NPs in German require the presence of some strongly case-marked element (determiner or adjective) in addition to the head noun in order to be acceptable (Gallmann & Lindauer, 1994; Schachtl, 1989; see also Eisenberg, 2013, 160).

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

2.1.2 Syntax of the alternating constructions

In order to understand the MNP alternation, it is vital to consider how 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 so-called strong and weak inflection patterns of German adjectives needs to be taken into account. In NPs with a strongly inflected determiner, attributive adjectives inflect according to the massively syncretic *weak* pattern. If there is no determiner (as is the case in the alternating constructions), attributive adjectives inflect like determiners themselves. This is called the *strong* inflectional pattern. For example, in the dative (governed here by the preposition *mit*), the strong suffix *-em* is on the

determiner in (4a) and the adjective bears the weak suffix *-en*. In (4b), the strong suffix *-em* appears on the adjective because there is no determiner.[7]

(4) a. mit ein-em stark-en Kaffee
with a strong coffee

b. mit stark-em Kaffee
with strong coffee

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 interpretation of the pattern shown in Table 1. If the adjective is classified as a determiner by virtue of carrying the inflectional markers which are otherwise characteristic of determiners, the PGC_{adj} is the appropriate choice. However, if it is classified as an adjective, the NAC_{adj} suggests itself because of the lack of a determiner.[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 explains why the alternation arises in the first place. In Section 2.2, I will argue that the NAC and the PGC constructions also have semantically distinguishable prototypes, and that the alternation between PGC_{adj} and NAC_{adj} is an alternation between these prototypes.

2.2 What controls the MNP alternation?

2.2.1 Prototype effects

My analysis of what controls the MNP alternation is based first and foremost on the idea that the PGC prototype expresses a pseudo-partitive where two discernible entities – the measure and the substance – are referenced. The NAC prototype, on the other hand, merely expresses a quantity, and the measure is not referenced as an entity in its own right. Prototypical exemplars are given in (5a) for the PGC and (5b) for the NAC.

7 In the masculine and neuter singular genitive, the strong and weak forms are indistinguishable. Since the alternation itself does not occur in the genitive (see Section 2.1), this does not create any complications for the present study.

8 The generative analysis presented in Bhatt (1990) comes close to this interpretation by analysing the kind NP in the PGC as a DP and in the NAC as an NP.

(5) a. Sie nahmen [einen Löffel [irgendeiner Medizin]$_{\text{Gen}}$]$_{\text{Acc}}$.
they took a spoon some medication
They took a spoon(ful) of some medication.

b. Sie kauften [drei Liter [Öl]$_{\text{Acc/caseless}}$]$_{\text{Acc}}$.
they bought three liters oil
They bought three liters of oil.

While the PGC$_{\text{det}}$ in (5a) allows an interpretation where speakers conceptualise the substance itself (i. e., the medication) and an actual spoon used to measure the quantity of the medication, the NAC$_{\text{bare}}$ in (5b) does not allow such an interpretation. While in the given examples, the effect is strongly supported by the choice of the measure noun lemmas, I argue that the meaning of the prototypes is independent of this. The independent meanings of the prototypes are the result of diachronic developments and grammaticalisation processes. Furthermore, I argue that the prototypical meanings are reflected in their usage patterns.

I begin by showing how the two meanings of the constructions emerge as a consequence of grammaticalisation processes of partitives and similar constructions. It is often assumed that pseudo-partitives and quantity constructions 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 path described by Koptjevskaja-Tamm (2001, esp. 526–530) can start out (in some languages) with constructions involving two referential nouns (not even necessarily forming a single and contiguous NP) and a *separative* meaning as in *(cut) two slices from the cake* (Koptjevskaja-Tamm, 2001, 535). In this type of construction, it is most obvious that two separate referents (in the given example the cake and the slices) are conceptualised. 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, especially when words like *bite* are no longer necessarily interpreted as a piece literally 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 two (now clearly distinct) available constructions have emerged diachronically from a single source through a complex reanalysis process, and the PGC is clearly the older construction (Zimmer, 2015, 2–4). As predicted by the grammaticalisation pattern just described, it still has the potential to form a true partitive (if the kind noun is definite). Conversely, the NAC lacks this ability to form true partitives and has gone further down the grammaticalisation path. It is thus not surprising that it has lost (at least prototypically) the semantics which allows both the measure

and the substance to be conceptually accessible as independent referents. As a consequence, we should expect a tendency for the NAC constructions to host more strongly grammaticalised measure nouns. For example, highly grammaticalised non-referential nouns like *Gramm* 'gram' and *Meter* 'metre' should occur proportionally more often in NAC constructions than in PGC constructions. If such preferences can actually be shown in usage patterns, it would lend strong support for the hypothesised difference in the meanings of the prototypes. In the corpus study, measure lemmas will therefore be annotated with appropriate semantic class labels to check whether semantic classes of measure nouns have different affinities to the two variants. The actual classification used is based on the list in Koptjevskaja-Tamm (2001, 530), but due to the low token frequencies in many of the potential classes, a very coarse classification was used in the end. With typical examples, the classes are: *Physical* (typically non-referential precisely measurable units such as *Liter* 'litre', *Meter* 'metre', *Gramm* 'gram'), *Container* (*Tasse* 'cup'), *Amount* (*Menge* 'amount'), *Portion* (natural portions like *Happen* 'bite' or *Krümel* 'crumb'). Notice that *Container* is the class containing words like *spoon* or *cup*, which often develop into partially grammaticalised physical measure nouns. The lemmas that did not fit into either of these classes were labelled *Rest*.

A second preference should also be observable as a consequence of the different meanings. 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 typical of quantities to be specified with numerical precision. This is most obviously so with the highly grammaticalised physical measure nouns like *centilitre*, which are very typically used with numerals instead of unspecific quantifiers, although both options are available in principle (*three centilitres* vs. *several centilitres*). 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, (6) shows the expected alternants under this hypothesis.

(6) 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 (6a), the measure noun is modified by a cardinal *drei* 'three', and hence the $\mathrm{NAC}_{\mathrm{adj}}$ is preferred. In (6b), the measure noun is modified by a non-cardinal determiner *einige* 'some', and the $\mathrm{PGC}_{\mathrm{adj}}$ is preferred. Especially with exact physical measure nouns (like *centilitre*), exact numerical quantification is invited. By hypothesis, however, this goes beyond a selection effect tied to measure lemmas, and cardinal quantifiers are expected to co-occur relatively more often with the $\mathrm{NAC}_{\mathrm{adj}}$.

Finally, it was found that the $\mathrm{PGC}_{\mathrm{adj}}$ is more typical of higher stylistic levels (edited, closer to the non-regional standard, more formal) and/or even exclusive to written language (see Hentschel, 1993, 320–323). The genitive – an intrinsic part of the PGC – is rarer in colloquial vernacular variants of German compared to the written standard. This is the result of a diachronic process wherein some (but by no means all) uses of the genitive are being replaced by other cases or periphrastic constructions (Fleischer & Schallert, 2011). Under an integral view of prototypes, which incorporates effects related to larger contexts and styles, such preferences can be part of what defines the construction prototypes, and the $\mathrm{PGC}_{\mathrm{adj}}$ should occur proportionally more often in more elaborate styles closer to the standard.

2.2.2 Exemplar effects

While the prototypes for the PGC and the NAC were specified with reference to high-level features in Section 2.2.1, at least one exemplar-driven similarity effect can also be expected to influence the MNP alternation. The measure and kind noun lemmas which occur in the alternating constructions obviously also occur in the non-alternating constructions. The relative frequency with which they occur in these stable and highly frequent cases – where choosing an alternative is impossible – could thus be a factor influencing the alternating, less stable case. Should this be confirmed, it would be highly implausible to conceive of such an effect as a prototype effect. In the corpus study reported in Section 3, a measure quantifying this influence will therefore be included as the *attraction strength*.

It should be noted that such an exemplar-type effect would have to be confirmed *in addition* to the predicted semantic prototype effect for measure lemmas described in Section 2.2.1, namely the effect of semantic classes of measure nouns. It must also be ensured that these influences at the lemma level are not spurious and just artefacts of mere lemma idiosyncrasies. This leads to a rather complex and truly multilevel model structure for the generalised linear mixed

models (GLMMs) to be used in Section 3.[9] While true multilevel models are not used very often in corpus linguistics – Gries (2015b) even calls the simpler varying-intercepts and varying-slopes models "underused" – they provide an excellent tool to describe situations where preferences at the sentence level and at lexical levels need to be integrated. The models do not differentiate in and of themselves between abstraction and exemplar effects, but they allow researchers to tune the degree of complexity of models, incorporating both types of effects according to their analysis and the phenomenon at hand. In Sections 3.2 and 3.3, multilevel models will therefore be used.

9 See Gelman & Hill (2006) (especially part 2) for a comprehensive text book with a focus on multilevel models.

3 Corpus study

3.1 Preliminiaries

3.1.1 Corpus choice

For the present study, I used the German *Corpus from the Web* (COW) in its 2014 version DECOW14A (Schäfer & Bildhauer, 2012, and Schäfer, 2015, as well as Biemann et al., 2013, and Schäfer & Bildhauer, 2013, for overviews of web corpora in general and the methodology of their construction), which contains almost 21 billion tokens.[10] I chose this corpus for two main reasons. First, the external validity of any study is increased through a higher heterogeneity of the sample (Maxwell & Delaney, 2004, 30), and the DECOW14A corpus has clearly a much more heterogeneous composition compared to the only other very large corpus of German, the DeReKo of the Institute for the German Language (Kupietz et al., 2010), 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 proofreaders who must adhere to normative guidelines might favour one alternative or the other explicitly. Web corpora, on the other hand, contain a significant amount of non-standard language from forums and similar sources. For such reasons, COW corpora have been used in a number of peer-reviewed publications, for example Goethem & Hiligsmann (2014), Goethem & Hüning (2015), Müller (2014), Schäfer (2016), Schäfer & Sayatz (2014), Schäfer & Sayatz (2016), and Zimmer (2015). Therefore, DECOW14A is a valid choice for this study.

10 The COW corpora (Dutch, English, French, German, Spanish, Swedish) are made available for free at https://www.webcorpora.org. At the time of this writing, a newer 2016 version DECOW16A has already been released.

11 It was shown in Bildhauer & Schäfer (2016) that, for example, the range of topics is much smaller in DeReKo compared to DECOW14A.

3.1.2 Bootstrapping pairs of lemmas

Among the factors potentially influencing the alternation (see Section 2) are lemma-specific effects. Therefore, it was 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 two-stage process in order to obtain a list of co-occurring measure nouns and kind nouns.

In the first step, I exported a list of all nouns from the DECOW14A01 subcorpus (one billion tokens large). Then, the one hundred most frequent mass nouns were extracted manually from this list. Mass nouns were defined as concrete nouns which genuinely denote a substance, combine with uninflected mass quantifiers such as *viel* 'much' and *wenig* 'little' (*viel Bier* 'much beer'), and form only sortal and unit plurals (such as the plural *Biere* 'types of beer' or 'glasses of beer'). Abstract nouns which partially behave like mass nouns (like *Spaß* 'fun' or *Gefahr* 'danger') were excluded because they are not measured in the same way as concrete mass nouns.

This list of mass nouns was used in the second step to derive a list of measure nouns co-occurring with the mass nouns. In order to generate this list, I utilised the fact that a direct sequence of two nouns almost always instantiates the bare-noun NAC if the second noun is a mass noun. I searched for all sequences N_1N_2 where N_2 was one of the mass noun lemmas extracted in the first step. 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 measure and mass nouns.

3.1.3 Variables and annotation

The full set of manually annotated variables for the main study is given in Table 2.[12] Notice first that *Construction* is the response variable with the values *PGCadj* and *NACadj*.

The variables *Kindattraction* and *Measureattraction* encode the ratio with which a given kind noun lemma or measure noun lemma occurs in the PGC_{det}

12 All numeric variables were z-transformed to facilitate their interpretation in the regression models.

Unit of reference	Variable	Type	Levels (for factors only)
Document	Badness	numeric	
	Genitives	numeric	
Sentence	Cardinal	factor	Yes, No
	Construction (response)	factor	NACadj, PGCadj
	Measurecase	factor	Nom, Acc, Dat
Kind lemma	Kindattraction	numeric	
	(Kindcollo)	numeric	
	Kindfreq	numeric	
	Kindgender	factor	Masc, Neut, Fem
Measure lemma	Measureattraction	numeric	
	(Measurecollo)	numeric	
	Measureclass	factor	Physical, Container, Amount, Portion, Rest
	Measurefreq	numeric	

Table 2: Annotated variables for the corpus studies

and the $\mathrm{NAC}_{\mathrm{bare}}$. They were calculated from auxiliary samples to be described in Section 3.3.1 as a log-transformed quotient. The higher the value for some noun, the (relatively) more often this noun occurs in the $\mathrm{PGC}_{\mathrm{det}}$. It could be argued that other measures of attraction strength should be used, for example those popularised in collostructional analysis (CA; Stefanowitsch & Gries, 2003; Gries & Stefanowitsch, 2004; see also Gries, 2015c). However, the goal here is to quantify how often lemmas occur in the $\mathrm{PGC}_{\mathrm{det}}$ and the $\mathrm{NAC}_{\mathrm{bare}}$, and these constructions do not compete at all but are rather mutually exclusive. While this does not preclude the use of CA, it is an open question whether the marginals are cognitively relevant in this case. After all, the main difference between the quotient used here and measures used in CA (originally signed logarithmised p-values from a Fisher test) is that they take the marginals into account. However, using Fisher-p-based collexeme strength instead of the raw frequency quotient was tried as an alternative (variables *Kindcollo* and *Measurecollo*). See Section 3.3.2 for a discussion of the negative results.

In Section 2.2, it was hypothesised that classes of measure lemmas might have different preferences for the two alternants. To capture this, class information was annotated for measure lemmas as *Measureclass*. The variable *Cardinal* encodes whether the MNP is modified by a cardinal.

To capture the influence of style mentioned in Section 2.2, two proxy variables were used. At the document level, the DECOW14A 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

language model of standard German. The paper also shows that the Badness score corresponds robustly with human raters' intuitions about text coherence and text quality. Documents with higher Badness usually contain more incoherent language and shorter sentences. If the $\mathrm{PGC}_{\mathrm{adj}}$ actually favours more elaborate stylistic levels, a high *Badness* should be correlated with fewer occurrences. Documents in DECOW14 have also been annotated with a variable called *Genitives*. The higher the values of this variable, the lower the proportion of genitives among all case-bearing forms is. A high number of genitives is also indicative of a more formal, elaborate style close to the written standard. However, since the $\mathrm{PGC}_{\mathrm{adj}}$ contains a genitive itself, the regressor variable *Genitive* and the document-level variable *Genitives* are not fully independent. 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.

Furthermore, to control for simple frequency effects, *Kindfreq* and *Measurefreq* are included as the logarithm-transformed frequencies per 1,000,000 words of each lemma, extracted from the frequency lists distributed by the DECOW14A corpus creators on their web page.

Finally, two variables were added as nuisance variables in the context of the present study. First, 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 more strongly against *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. The case of the measure noun was manually annotated (variable *Measurecase*). Second, due to differences in case syncretisms, it is likely that feminine kind nouns have slightly different preferences than masculine and neuter ones, and hence the appropriate variable *Kindgender* was annotated.

3.2 Pre-study: main prototype effects in the non-alternating constructions

The main study to be reported below deals exclusively with the alternating constructions, as is customary in alternation modelling. However, the prototypical features described in Section 2.2.1 should be observable in the non-alternating cases as well. Therefore, in this section I examine the distribution of the prototypical features in the non-alternating cases to see whether they are in accord with the theoretical predictions.

Model level	Regressor	p_{PB}	Factor level	Coefficient	CI low	CI high	CI excludes 0
1	Badness	0.042		0.120	-0.002	0.228	
	Cardinal	0.001	No	1.419	0.869	1.993	*
	Genitives	0.001		-0.710	-0.815	-0.556	*
2 (Kindlemma)	(Kindfreq)	(0.781)					
	(Kindgender)	(0.199)					
2 (Measurelemma)	Measureclass	0.005	Container	0.445	-0.579	1.532	
			Rest	1.745	0.649	2.831	*
			Amount	1.597	0.125	2.751	*
			Portion	1.782	0.771	2.690	*
	(Measurefreq)	(0.265)					

Table 3: Coefficient table with 95% bootstrap confidence intervals for the pre-study; the intercept is -5.370

3.2.1 Sampling and annotation

For this pre-study, each of the 2,365 combinations of measure noun lemma and kind noun lemma were queried in the $\mathrm{NAC}_{\mathrm{bare}}$ and the $\mathrm{PGC}_{\mathrm{det}}$ with a limit of 100 randomly chosen results for each single query. From the resulting 35,766 sentences, 5,000 were randomly sampled for the statistical analysis. All features described in Section 3.1.3 were annotated except for the ones which do not apply in the non-alternating case (*Kindattraction*, *Measureattraction*, and *Measurecase*).

3.2.2 Statistical model

The annotated concordance was analysed in the form of a multilevel logistic regression model using R (R Core Team, 2014) and the *lme4* package (Bates, 2010; Bates et al., 2015b). An alternative *BOBYQA* optimiser from the *nloptr* package (Johnson, 2017) was used for all fits with *lme4* reported in this paper.

The coefficient estimates are specified in Table 3 for each regressor (column *Coefficient*). Given the coding of the response variable, coefficients leaning to the positive side can be interpreted as characterising a configuration typical of the $\mathrm{PGC}_{\mathrm{det}}$. For a robust quantification of the precision of the estimation, I ran a parametric bootstrap (using the *confint.merMod* function from *lme4*) with 1,000 replications and using the percentile method. The resulting 95% bootstrap confidence intervals are reported in Table 3 (columns *CI low* and *CI high*). The column *CI excludes 0* shows an asterisk for those intervals that do not include 0. Furthermore, for each regressor, a p-value was obtained by dropping the regressor from the full model, re-estimating the nested model, and comparing it

to the full model. Instead of inexact Wald approximations and Likelihood Ratio Tests, I used a drop-in bootstrap replacement for the Likelihood Ratio Test in the form of the function *PBmodcomp* from the *pbkrtest* package (Halekoh & Højsgaard, 2014). I call the corresponding value p_{PB} (column p_{PB}). Regressors which did not reach sig=0.05 in in the PBmodcomp test (*Kindfreq*, *Kindgender*, and *Measurefreq*) were removed from the model, appear in brackets in Table 3, and have no coefficient estimates.

The overall intercept is -5.370, and it represents *Cardinal=Yes*, *Measureclass=Physical* and 0 for all (z-transformed) numeric regressors. The standard deviation of the random intercepts is 1.157 for *Measurelemma* ($p_{\mathrm{PB}} = 0.001$) and 0.923 for *Kindlemma* ($p_{\mathrm{PB}} = 0.002$). Nakagawa & Schielzeth's pseudo-coefficients of determination are $R^2_m = 0.278$ and $R^2_c = 0.566$ (Gries, 2015b; Nakagawa & Schielzeth, 2013).

3.2.3 Interpretation

The coefficients of determination indicate that the influence of the fixed effects (marginal R^2) is adequately strong, and that lemma-specific effects are equally strong with a difference between the two coefficients of 0.289.

The main prototypical factors *Cardinal* and *Measureclass* both pass the PBmodcomp test at sig=0.05 and play the expected role. Non-referential physical measure nouns (such as *Gramm* 'gram' or *Liter* 'litre') with a high degree of grammaticalisation favour the $\mathrm{NAC}_{\mathrm{bare}}$. At the other end of the scale, nouns denoting natural portions like *Haufen* 'heap', *Bündel* 'bundle', *Schluck* 'gulp' favour the $\mathrm{PGC}_{\mathrm{det}}$. Also, The presence of a cardinal modifier clearly favours the $\mathrm{NAC}_{\mathrm{bare}}$. The stylistic factor *Genitives* shows the predicted influence inasmuch as a general lack of genitives (which is encoded as a *higher* value of the variable) favours the $\mathrm{NAC}_{\mathrm{bare}}$. *Badness* points into the wrong direction, but it barely reaches sig=0.05, and its confidence interval includes 0. The lemma frequency regressors both clearly fail the PBmodcomp test. I will return to these results in Section 3.3.3.

3.3 Main study: the alternation

3.3.1 Sampling and annotation

For the main study, each of the 2,365 noun–noun combinations was queried in the alternating constructions $\mathrm{PGC}_{\mathrm{adj}}$ and $\mathrm{NAC}_{\mathrm{adj}}$ using a sub-corpus of

DECOW14A (slices DECOW14A01 to DECOW14A10 with a total size of 10 billion tokens). For further manual annotation, a random sub-sample was generated. For each mass noun, the concordance was downsampled randomly to maximally 100 sentences, which resulted in 6,843 sentences. In the manual annotation, the sample was further reduced to 5,063 sentences 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 $\mathrm{PGC}_{\mathrm{adj}}$.[13]

Finally, two auxiliary samples were also drawn. As mentioned in Section 2.2, the distribution of the measure 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 process described above 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 $\mathrm{NAC}_{\mathrm{bare}}$.

3.3.2 Statistical model

A multilevel logistic regression model was fit which models the influence of the regressors specified in Table 2 on the probability that the $\mathrm{PGC}_{\mathrm{adj}}$ is chosen over the $\mathrm{NAC}_{\mathrm{adj}}$. All regressors from Table 2 were included, and the measure lemma and the kind noun lemma were specified as varying-intercept random effects. The sample size was n=5,063 with 1,134 cases of $\mathrm{PGC}_{\mathrm{adj}}$ and 3,929 cases of $\mathrm{NAC}_{\mathrm{adj}}$. The results of the estimation are shown in Figure 1 and in Table 4. The intercept comprises *Cardinal=Yes*, *Measurecase=Nom*, *Kindgender=Masc*, *Measureclass=Physical*, and 0 for all numeric z-transformed regressors. It was estimated at -3.548.

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

13 In a similar fashion, the 100 most frequent measure nouns occurring with plural kind nouns were listed and queried, resulting in a sample of 871 sentences. As stated in Section 2.1.1, 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 with the data package accompanying this paper.

Standard diagnostics show that the model quality is quite good. Nakagawa & Schielzeth's pseudo-coefficients of determination are $R^2_m = 0.409$ and $R^2_c = 0.495$. The rate of correct predictions is 0.843, which means a proportional reduction of error of $\lambda = 0.297$. Generalised variance inflation factors for the regressors were calculated to check for multicollinearity (Fox & Monette, 1992; Zuur et al., 2010), and the highest corrected $\text{GVIF}^{1/2\text{df}}$ was 1.520 for *Cardinal*. The lemma intercepts have standard deviations of $\sigma_{\text{Measurelemma}} = 0.448$ and $\sigma_{\text{Kindlemma}} = 0.604$. Only *Kindfreq* ($p_{\text{PB}} = 0.095$) could be seen as slightly too high to be convincing, failing at sig=0.05.

Using signed logarithmised Fisher p-values as a measure of collexeme strength (*Measurecollo* and *Kindcollo*) instead of the quotient for the attraction strength (*Measureattraction* and *Kindattraction*) (see Section 3.1.3) was not successful. The p_{PB} value for *Kindcollo* was 0.191 and the one for *Measurec-*

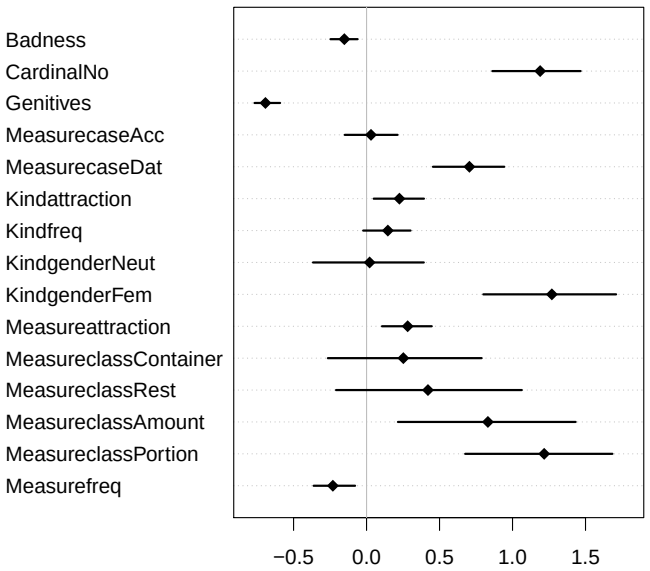

Fig. 1: Coefficients with 95% confidence intervals (for details see text); the intercept is -3.548

Model level	Regressor	p_{PB}	Factor level	Coefficient	CI low	CI high	CI excludes 0
1	Badness	0.002		-0.152	-0.247	-0.061	*
	Cardinal	0.001	No	1.189	0.862	1.466	*
	Genitives	0.001		-0.693	-0.768	-0.592	*
	Measurecase	0.001	Acc	0.030	-0.150	0.212	
			Dat	0.705	0.455	0.944	*
2 (Kindlemma)	Kindattraction	0.020		0.225	0.049	0.393	*
	Kindfreq	0.095		0.146	-0.023	0.301	
	Kindgender	0.001	Neut	0.021	-0.367	0.392	
			Fem	1.269	0.800	1.709	*
2 (Measurelemma)	Measureattraction	0.001		0.282	0.106	0.447	*
	Measureclass	0.001	Container	0.252	-0.265	0.788	
			Rest	0.421	-0.209	1.063	
			Amount	0.831	0.215	1.432	*
			Portion	1.217	0.675	1.684	*
	Measurefreq	0.005		-0.231	-0.363	-0.079	*

Table 4: Coefficient table with 95% bootstrap confidence intervals for the main study; the intercept is -3.548

ollo was 0.443. The coefficients of determination dropped to $R^2_m = 0.376$ and $R^2_c = 0.480$.

3.3.3 Interpretation

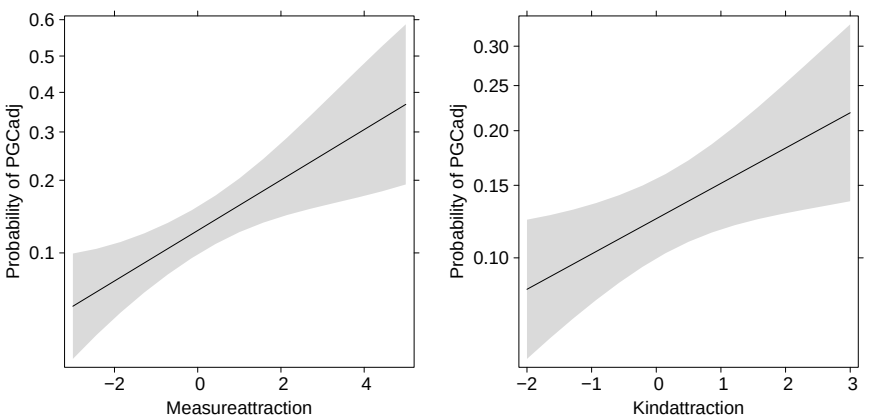

Fig. 2: Effect plots for the regressors *Measureattraction* and *Kindattraction*; y-axes are not aligned

The results generally confirm the hypotheses from Section 2.2. First, the exemplar effect (the influence of the non-alternating $\mathrm{PGC}_{\mathrm{det}}$ and $\mathrm{NAC}_{\mathrm{bare}}$) was shown (see the effect plots in Figure 2).[14] The effect is as expected: if a lemma appears relatively more often in the $\mathrm{PGC}_{\mathrm{det}}$ (compared to its frequency in the $\mathrm{NAC}_{\mathrm{bare}}$), the $\mathrm{PGC}_{\mathrm{adj}}$ tends to be chosen over the $\mathrm{NAC}_{\mathrm{adj}}$ with this specific lemma. The effect for measure nouns is stronger, and it was estimated with higher precision.

An interesting picture emerges for the lemma frequencies. A higher-than-average lemma frequency of measure nouns favours the $\mathrm{NAC}_{\mathrm{adj}}$, which is as expected if we assume at least a tendency for highly grammaticalised words to be more frequent. With kind nouns, higher frequency seems to favour the $\mathrm{PGC}_{\mathrm{adj}}$. However, this effect has no clear theoretical interpretation, and its coefficient estimate is imprecise (not significant at sig=0.05). The effect can therefore be ignored or treated as a nuisance variable.

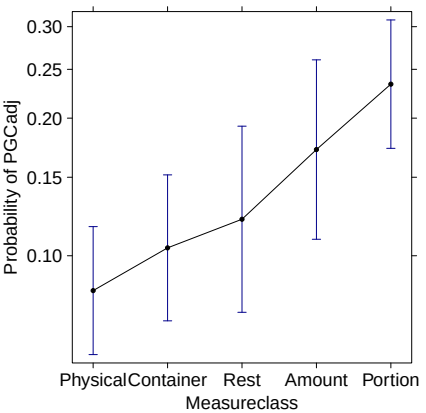

Fig. 3: Effect plot for the regressor *Measureclass*

In Section 2.2, it was also hypothesised that classes of measure nouns with a higher degree of grammaticalisation should favour the $\mathrm{NAC}_{\mathrm{adj}}$. The *Measureclass* second-level predictor reaches sig=0.05 in the PBmodcomp test. 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

14 Effect plots were created using the *effects* package (Fox, 2003).

denoting natural portions like *Haufen* 'heap', *Bündel* 'bundle', *Schluck* 'gulp' favour the $\mathrm{PGC}_{\mathrm{adj}}$. These are referential nouns, confirming the hypothesis that it is prototypical of the PGC to contain two referential nouns, while the NAC prototypically only contains one (the kind noun).

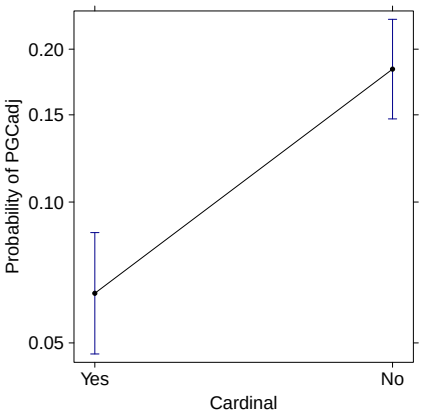

Fig. 4: Effect plot for the regressor *Cardinal*

Figure 4 shows that cardinals indeed influence the choice of the alternant, and that cardinals have a strong tendency to co-occur with the $\mathrm{NAC}_{\mathrm{adj}}$.

The style-related proxy variables point in the expected direction. Increased *Badness* of the document favours the $\mathrm{NAC}_{\mathrm{adj}}$, and so does a lower density of genitives. While these are merely proxies to style, this result can at least encourage future work into stylistic effects.

The influence of *Measurecase* is as predicted in previous analyses. A measure noun in the dative favours the $\mathrm{PGC}_{\mathrm{adj}}$ (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.

To close this section, I now compare the results of the pre-study and the main study. Even though the non-alternating constructions are surely subject to additional constraints, the coefficients align neatly in many cases. First of all, the intercepts encode a similar overall dominance of the NAC constructions (pre-study: -5.370; main study: -3.548). For the *Cardinal* effect, the coefficients 1.419 (pre-study) and 1.189 (main study) have the same sign and magnitude and mostly overlapping confidence intervals. The levels of *Measureclass* have comparable coefficients, although the divergence is larger. In both studies, the

non-referential measure nouns in the *Physical* class (on the intercept) are most clearly associated with the NAC constructions. Also, in both cases, the *Container* class is closest to *Physical* with the same sign and magnitude (pre-study: 0.445; main study: 0.252). The main difference – if we ignore the *Rest* class which can be expected to show no clear tendency – is that *Amount* and *Portion* are slightly closer together in their tendency to favour the PGC constructions in the pre-study (*Amount* 1.597 and *Portion* 1.782 in the pre-study vs. *Amount* 0.831 and *Portion* 1.217 in the main study). The difference is not huge, and the overall order of the coefficients is the same. The *Genitives* effect also converges with -0.710 in the pre-study and -0.693 in the main study. The two studies do not converge with respect to the *Badness* variable, which is not significant in the pre-study. However, it would be surprising if we achieved perfectly converging results given that even real effects are missed at certain rates in empirical studies. In the case of *Badness*, we see that even in the main study, the post-hoc effect size is small (-0.152), and thus even at the impressive sample size of roughly 5,000 in both studies, the effect might simply be too weak to be detected reliably. Finally, the frequency effect *Measurefreq* is detected in the main study but not in the pre-study, while the *Kindfreq* effect is essentially absent in both studies. In connection with this, it is revealing to look at the coefficients of determination. In the pre-study, a much greater proportion of the variance is explained by taking the lemma random effects into account ($R^2_m = 0.278$ and $R^2_c = 0.566$) than it is in the main study ($R^2_m = 0.376$ and $R^2_c = 0.480$). Thus, the frequency effect for measure lemmas might be swamped by the lemma random intercepts. All things considered, the studies have shown that the predicted prototype and exemplar effects are reflected in usage data for both the alternating and the non-alternating constructions.

4 Experiments

4.1 Experiment 1: forced-choice

In this section, the probabilities of the alternants (as calculated from attested material in the corpus study) are correlated with the reactions of participants in controlled experiments. Thus, a direct link is established between output material found in corpora and the behaviour of linguistic agents.

The decision to use attested material and perform a global validation of the corpus-based model was not arbitrary. Alternatively, one could have created stimuli where the features favouring the alternatives were permuted and thus tested directly. The answer is related to what was said in Section 1 about the rich multi-factorial data sets used in corpus studies and the comparatively restricted ones in experiments. Looking at Table 2, we see one binary factor (*Cardinal*), two three-level factors (*Measurecase* and *Kindgender*) and one five-level factor (*Measureclass*). Since the dependent variable is binary (*Construction*), permuting these factors alone leads to 180 different possible permutations. Additionally, the numeric variables *Kindattraction* and *Measureattraction* would also have to be tested at several values. It is obvious that this is impossible in a controlled experiment, given that a participant can be exposed to no more than roughly one hundred sentences, most of which would be fillers rather than target sentences. Thus, while the present approach only allows for a global test of the corpus-derived model, this appears to me to be the only feasible way.[15] It is, in the sense of Section 1, an optimal synthesis of data-rich multifactorial corpus studies and experimental validation, and it is one that has been used at least since the seminal Bresnan et al. (2007) paper – although rarely, as Divjak et al. (2016b, 3–4) point out.

4.1.1 Setup, stimuli, and participants

The first experiment tests preferences for constructions explicitly in a forced-choice task. Participants had to choose between two sentences that differed only in that one contained the $\mathrm{NAC}_{\mathrm{adj}}$ and the other one contained the $\mathrm{PGC}_{\mathrm{adj}}$. In the analysis, the probabilities assigned to the stimuli by the corpus-derived model were compared to the frequency with which participants chose the alter-

15 For the same reasons, the experimental data are also much too sparse to perform post-hoc analyses with respect to the single regressors in a sound way.

	Masculine/Neuter	Feminine
high prob. for PGC$_{adj}$	4 sentences	4 sentences
low prob. for PGC$_{adj}$	4 sentences	4 sentences

Table 5: The four groups of sentences chosen as stimuli; in each group of four sentences, combinations of important factor values were made unique whenever possible

nants. There participants were 24 native speakers of German without reading or writing disabilities, aged 19 to 30, and living permanently in Berlin. They were recruited from introductory linguistics courses at Freie Universität Berlin. Although the experiment was conducted in the last four weeks of their first semester, participants had no deeper 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 simplified to avoid influences from contextual nuisance factors as much as possible. 16 MNPs were sampled from the concordance, making sure that any simplifications and normalisations did not affect any of the regressors used in the corpus study. In the simplified sentences, the case, number, etc. of the MNP remained the same as in the attested sentence, as did the choice of lexical material within the MNP. Eight sentences contained masculine or neuter kind nouns, and the other eight contained feminine kind nouns. Furthermore, in each of the masculine/neuter and feminine groups, four sentences originally containing the $\mathrm{NAC}_{\mathrm{adj}}$ and four sentences originally containing the $\mathrm{PGC}_{\mathrm{adj}}$ were chosen. Moreover, the sentences were sampled as typical examples of $\mathrm{PGC}_{\mathrm{adj}}$ (high probability assigned by GLMM) and $\mathrm{NAC}_{\mathrm{adj}}$ (low probability assigned by GLMM), respectively.[16] High and low probabilities were defined (roughly) 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. For calculating the model predictions, the document-level variables *Badness* and *Genitives* were set to 0, which is the mean for z-transformed variables. The design is summarised in Table 5.

The final pairs of stimuli were the sentence containing the attested and preferred alternant (according to the corpus GLMM) on the one hand and a modified version containing the dispreferred alternant on the other hand. They

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

were presented next to each other, and a 20 second time limit for each choice was set.[17] The order of sentences was randomised for each participant, and the position of the alternants on the screen (left/right) was randomised per participant and sentence. As fillers, 23 pairs of sentences exemplifying similar but unrelated alternation phenomena from German morpho-syntax were used. Thus, participants saw 39 pairs of sentences and 78 sentences in total. 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).

4.1.2 Statistical model

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

A good amount of the variance can be accounted for by idiosyncrasies of single sentences ($\sigma_{\mathrm{Item}} = 1.217$). Also, among participants, there are clearly different preferences ($\sigma_{\mathrm{Participant}} = 0.412$). On the extreme ends, one participant chose the $\mathrm{PGC}_{\mathrm{adj}}$ in 13 of 16 cases, and two participants only chose it in 5 of 16 cases. The regressor *Modelprediction* achieves $p_{\mathrm{PB}} = 0.003$ (1,000 replications) and is estimated at 4.389 relative to an intercept of -1.270. The confidence interval from a parametric bootstrap (1,000 replications, percentile method) for the regressor is acceptable but a tad large with a lower bound of 1.788 and an upper bound of 6.599. The pseudo-coefficients of determination are $R^2_m = 0.185$ and $R^2_c = 0.455$, which means that roughly 19% of the variance in the data can be explained by considering only the predictions from the corpus-based GLMM.

4.1.3 Interpretation

The marginal R^2 indicates a weak result for the fixed effects part of the model, which is nonetheless worthy of mention (close to 0.2). The effect display for the single fixed regressor *Modelprediction* is given in Figure 5. The higher the prob-

17 No participant ever exceeded the time limit.

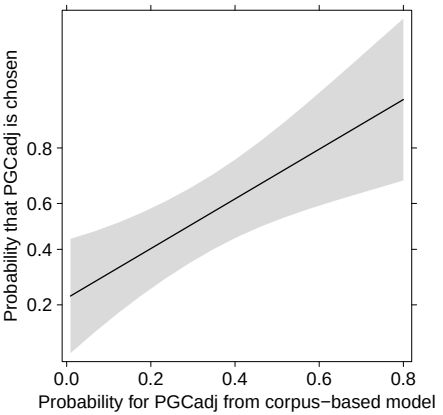

Fig. 5: Effect plot for the multilevel logistic regression in the forced-choice experiment: predictability of participants' choices using the probabilities derived from the corpus-based GLMM

ability of the PGC_{adj} predicted from usage data, the more often participants chose the PGC_{adj} alternant in the forced-choice task. A closer look at the results in the form of the spineplot in Figure 6 shows, however, that it was likely an idiosyncracy of a single sentence which spoiled an otherwise much better correlation. The problematic sentence with a model prediction of 0.548 is given in (7) in the PGC_{adj} variant.

(7) Man machte mal wieder viel Lärm um [jede Menge [heißer
one made once again much noise about any amount hot
Luft]$_{\text{Gen}}$]$_{\text{Acc}}$.
air

People made much ado about nothing once again.

In retrospect, this stimulus was badly chosen because *heiße Luft* 'ado' (literally 'hot air') is a fixed metaphorical expression. This obviously influences the reactions of participants, and a revised and improved experiment might lead to a much better fit in future research.

In principle, random slope for *Modelprediction* varying by item could also remedy problems with individual stimuli at least partially. Therefore, a model with random slopes for *Modelprediction* varying by both random effects (*Participant* and *Item*) was specified and estimated. The random slope for participants was added to comply with Barr et al. (2013, 257) who predict "catastrophically high Type I error rates" for experimental designs with within-subject manip-

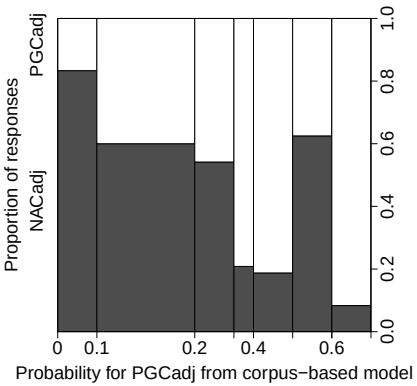

Fig. 6: Spineplot of the proportion of responses plotted against the predictions from the corpus-based model in the forced-choice experiment

ulations if random effects structures are not kept maximal. The coefficient of the fixed effect changed noticeably but not enough to change the interpretation (5.408 relative to an intercept of -1.304), and the marginal R^2_m rises to 0.213 ($R^2_c = 0.488$). In line with expectations, the standard deviation in the random slopes for *Item* is high at 5.996. However, the covariance parameters were estimated at exactly -1, which is a clear sign that the variance-covariance matrix could not be estimated successfully. The same was true for models with only an *Item* and a *Participant* random slope. This is exactly the kind of model overparametrisation criticised in Bates et al. (2015a) and Matuschek et al. (2017). The available data are insufficient to estimate the parameters of the more complex model with varying slopes.[18]

In summary, the forced-choice experiment succeeded in corroborating the results from the corpus study inasmuch as the preferences extracted from usage data correspond to native speakers' choices, but the correlation is weak, likely at least in part due to problems with individual test items and/or too sparse data.

18 Bates et al. (2015a, 1) state: "We show that failure to converge typically is not due to a suboptimal estimation algorithm, but is a consequence of attempting to fit a model that is too complex to be properly supported by the data, irrespective of whether estimation is based on maximum likelihood or on Bayesian hierarchical modelling with uninformative or weakly informative priors. Importantly, even under convergence, overparameterization may lead to uninterpretable models."

4.2 Experiment 2: self-paced reading

4.2.1 Setup, stimuli, and participants

The second experiment was conducted in a more implicit paradigm. It is expected that reading less typical alternants 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 presented. In a very similar fashion, Divjak et al. (2016a) 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.

The exact same stimuli as in the forced-choice experiments were used. Each participant read both the 16 sentences with the alternant predicted by the corpus model and the 16 modified sentences with the alternant that the corpus model did not predict.[19] To minimise repetition effects, the stimuli for each participant were separated into two blocks of 16 targets and 33 fillers per block. In the experiment, participants first read all sentences from the first block, then all sentences from the second block. From each target sentence pair, one sentence was assigned to the first block and the other sentence to the other block. The assignment of members of the individual sentence pairs to the blocks was randomised for each participant individually, as was the order within each block. The fillers also came in pairs such that the second block exclusively contained sentences to which participants had been exposed in the first block in slightly modified form. In total, each participant read 98 sentences. After each sentence, participants had to answer simple (non-metalinguistic) yes-no questions about the previous sentence as distractors. The distractor questions were different between the first and the second blocks. There were 38 participants recruited in exactly the same manner as for the experiment reported in Section 4.1. None of them had ever participated in any kind of reading experiment, and none of them took part in the first experiment. The experiment was conducted using *PsychoPy*.

The reading times were residualised per speaker based on the reading times of all words (not just the targets) for that speaker. The adjective and the kind noun (i. e., the constituents bearing the critical case markers) were used as the target region, for instance the bracketed words in the example *zwei Gläser* [*spru-*

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

Regressor	Coefficient	CI low	CI high	CI excludes 0
Construction=PGCadj	0.054	0.012	0.095	*
Modelprediction	-0.003	-0.113	0.110	
Position	-0.005	-0.005	0.004	
Construction=PGCadj:Modelprediction	-0.125	-0.234	-0.023	*

Table 6: Fixed effect coefficient table for the LMM used to analyse the self-paced reading experiment; the intercept is 0.829

delndes Wasser] 'two glasses of sparkling water'. Outliers farther than 2 interquartile ranges from the mean logarithmised residualised reading time were removed (64 data points), resulting in a total number of n=1,152 observations.

4.2.2 Statistical model

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 higher the GLMM prediction is, the more typical the sentence is for containing the $\mathrm{PGC}_{\mathrm{adj}}$. It is therefore expected that reading times are higher when the value of *Modelprediction* is higher but the sentence contains the $\mathrm{NAC}_{\mathrm{adj}}$. However, when the sentence contains the $\mathrm{PGC}_{\mathrm{adj}}$, reading times should be lower when *Modelprediction* is higher. To account for this, an interaction between *Modelprediction* and *Construction* (levels *PGCadj* and *NACadj*) 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. Random intercepts were specified for *Participant* and *Item* (the 16 sentence pairs are one *Item* each).[20]

Table 6 shows the coefficient estimates with a 95% parametric bootstrap confidence interval (1,000 replications, percentile method). The standard deviation of the participant intercepts is $\sigma_{\mathrm{Participant}} = 0.079$ and of the item intercepts $\sigma_{\mathrm{Item}} = 0.037$. Comparing the full model to a model without the main regressor *Modelprediction* (and consequently also without the interaction with

20 Again, all attempts to include random slopes resulted in the variance-covariance matrix not being properly estimated (1 or -1 covariance parameters).

Construction) in a PBmodcomp test gives $p_{\text{PB}} = 0.036$. The pseudo-coefficients of determination are $R^2_m = 0.239$ and $R^2_c = 0.346$.

An alternative Gaussian generalised additive model with an identity link was also fit (see Divjak et al., 2016a) using the *mgcv* package (Wood, 2011). The full results are included in the data package for this paper, but the fit was not better than with the LMM reported above. The estimated smoother for the *Modelprediction* variable is essentially linear, and the R^2 (corresponding to the marginal R^2 of the LMM) was 0.237.

4.2.3 Interpretation

The coefficients of determination indicate that there is a noteworthy correlation between the reactions of the subjects and the corpus-derived probabilities (marginal R^2) and that there is some between-subject variation: the difference between R^2_m and R^2_c is 0.107.

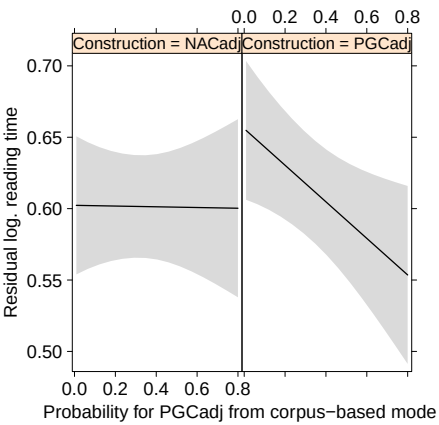

Fig. 7: Effect plot for the LMM in the self-paced reading experiment: modelling participants' residualised log reading times on the probabilities given by the corpus-based GLMM

The effect plot for the interaction of interest is shown in Figure 7. The estimate for the sentences with NAC_{adj} is obviously imprecise, and no significant 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 6) and from the PBmodcomp

test reported above. The PGC_{adj} brings about an increased reading time, which is plausible because it is the much rarer construction (see Section 3). However, if it occurs in a prototypical context and with typical lexical material, reading times drop. This can be seen in the downward slope in the right panel of Figure 7. This fits into the general picture inasmuch as the construction with the lower frequency might be developing towards a more sharply defined prototype.[21] Conversely, the NAC_{adj} (like the NAC in general) might be the highly frequent default which does not incur a reading time penalty, even if it is not the optimal choice in the given context and with the given lexical material.

In Section 5, I take stock and summarise the contributions of the present study to the research on alternations in cognitive linguistics.

21 In this context, it should be remembered from Section 3 that even the PGC_{det} is much rarer that the NAC_{bare} (17,252 vs. 315,635 occurrences in the auxiliary corpus samples).

5 Conclusions

The studies reported in Sections 3 and 4 have confirmed the theoretical model proposed in Section 2, which predicted prototype effects and exemplar effects to jointly determine which of the two alternating constructions is chosen. More specifically, the pre-study and the main study provided evidence that the NAC constructions favour more strongly grammaticalised measure nouns, modification by cardinals, and styles where the genitive is underrepresented and the text is less coherently and elaborately written. The overall convergence between the pre-study and the main study and the convergence between corpus data and the experimental cross-validation strengthen the validity of this study. Also, the fact that the result for *Measurecase* converged with previous research (Zimmer, 2015) lends credibility to the results.

Many relevant conclusions can be drawn based on the research presented here. Although direct inferences from corpus data to mental representations are problematic (see Section 1), the two attraction features (*Kindattraction* and *Measureattraction*) are virtually impossible to conceive of as prototype effects and thus favour an exemplar view or a mixed exemplar-prototype view. As discussed in Section 1, it is unnecessary to follow an extreme route, be it radical prototype theory or radical exemplar theory, given recent developments in cognitive science and cognitive linguistics. While it is surely an intentionally overstated comment, Kapatsinski (2014, 15) even suggests that "[i]n the extreme, some speakers' heads could host exemplar models, and some could contain fairly abstract grammars, and the produced output would be essentially identical". If this is the case (even to a less extreme degree), corpora only provide pooled data averaged over speaker grammars with varying levels of abstraction. Clearly, more research is required in this direction.

Another important aspect of the research presented here is the validation of the results derived from corpus data in two experimental paradigms. While such cross-validations have been done (with varying success, see Section 1) for over a decade, they have not yet become standard procedure. As Divjak et al. (2016b, 3–4) put it:

> There are now a number of published multivariate models that use data[,] extracted from corpora […] to predict the choice for one morpheme, lexeme or construction over another. However, […] only a small number of these corpus-based studies have been cross-validated […]. Of these cross-validated studies, few have directly evaluated the prediction accuracy of a complex, multivariate corpus-based model on humans using authentic corpus sentences […].

The question now arises whether convergence was reached in the present case or not. The answer is a clear yes. The predictions made by the corpus-based model were a significant factor, both in the more explicit paradigm (forced choice) and the implicit paradigm (self-paced reading). This shows that the model indeed partially predicts the variant which language users expect. The overall effect as measured in the R^2 was acceptable but not strong (roughly 0.2) in both cases, however. While this could be traced back at least partially to one suboptimally chosen stimulus, we really need to consider what kind of convergence we expect to see between corpus data and experiments. The main corpus study had $R^2_m = 0.409$ and $R^2_c = 0.495$, which is good but not anywhere near a perfect fit. With the added inter-speaker variability brought into play by the experimental setup (compared to the averaging across thousands of speakers in the corpus study), a perfect fit cannot be expected. Furthermore, like many alternations in German which are often labelled as *Zweifelsfälle* ('cases of doubt') in the traditional literature (Duden, 2011; Klein, 2009), the MNP alternation is a case where speakers often have no clear intuition, and a lot of free variation seems to be involved. Rarely do speakers feel that one alternative is clearly odd or ungrammatical. Additionally, cases of doubt involving the genitive are often a matter of fierce normative public debate, and especially the forced choice paradigm does not effectively prevent participants from making normative judgements. This might even account for the slightly better fit in the self-paced reading experiment, where normative considerations are suppressed. Thus, the present study shows that what counts as convergence between corpus and experimental results should be gauged considering the nature of the phenomenon at hand, the source of the corpus data, and the experimental paradigm. Clearly, more case studies using diverse and different corpora are needed, and it should become standard practice to cross-validate them using experimental methods. Given that a single failure to achieve convergence does not provide conclusive evidence for divergence, many more studies need to be published to the point that meta-analysis becomes possible.[22]

On a larger methodological scale, this paper also makes a number of contributions. The statistical model was a true multilevel model (see Section 3.3.2), demonstrating how multilevel modelling helps to specify complex hierarchies of abstract features, item-specific tendencies, and exemplar effects at the level of observations (sentences; first level) and lemmas (second level). Gries (2015b)

22 This is even more vital considering the variation in results from experimental work. See, for example, the impressive list of different reading time results from ten papers on Chinese relative clause processing in Vasishth (2015, 8).

still calls mixed models "underused" in corpus linguistics, and multilevel models are consequently also an underused tools. At the same time, the ill effects of overparametrisation of mixed models with varying slopes as criticised by Bates et al. (2015a) were demonstrated. Furthermore, fitting an additive model to the reading time data did not improve the fit as there were no non-linearities in the data. While Divjak et al. (2016a) also use attested sentences, they find that an additive model helped to deal with non-linearities. This is clearly another area where only more studies can lead to clarification. Finally, much like Dąbrowska (2014) found that speakers' knowledge of collocations was not matched by a set of standard measures of collocation strength extracted from corpora, a simple quotient based on raw frequencies for the attraction strength performed much better in the present study compared to collexeme strength in the form of logarithmised Fisher p-values. While this does not allow the conclusion that collexeme strength has no cognitive reality, it might indicate that for measuring attraction effects exerted by non-alternating constructions on alternating constructions, taking the marginals into account might not be crucial.[23]

Besides the methodological aspects mentioned above, future work could extend the validity of the results presented here through a more in-depth look at stylistic or register effects, for example using the new Biber-style annotations (Biber, 1988) which will be released with a new version of the DECOW corpus soon according to its creators. A reviewer also pointed out that strongly lexicalised adjective-noun combinations like *schwarzer Tee* 'black tea' might have a tendency to occur in the $\mathrm{NAC}_{\mathrm{adj}}$ because they have more compound-like qualities, blocking strong case inflection in between them. While this potential effect was impossible to incorporate into the present study, it could be integrated into future research on the phenomenon.

In closing, I want to point out that the so-called *case of doubt* in German morpho-syntax – like the measure noun phrase alternation – are in fact ideal test cases for probabilistic modelling of alternation phenomena in cognitive linguistics. Doing more research on them could help to provide answers to many of the fundamental and methodological issues raised here.

23 However, many different measures have been proposed within the collostructional framework (see Gries, 2015a for an overview of them and a defense of the framework). Future research might show significant differences between them for the problem at hand.

Acknowledgment: I thank (in alphabetical order) Felix Bildhauer, Susanne Flach, Elizabeth Pankratz, Samuel Reichert, Ulrike Sayatz, and Christian Zimmer for valuable discussions and comments. Also, I would like to thank the reviewers for Cognitive Linguistics as well as associate editor Dagmar Divjak for insightful comments which helped to improve the quality of this paper significantly. Furthermore, I thank Ulrike Sayatz for helping me to recruit the participants for the experiments. Elizabeth Pankratz thankfully also fixed my English. Finally, I am grateful to my student assistants Kim Maser for her work on the annotation of the concordances and Luise Rißmann for supervising most of the experiments. The research presented here was made possible in part through funding from the *Deutsche Forschungsgemeinschaft* (DFG, personal grant SCHA1916/1-1).

References

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

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, Reinhold Kliegl, Shravan Vasishth & R. Harald Baayen. 2015a. Parsimonious mixed models. https://arxiv.org/abs/1506.04967.

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

Bates, Douglas M. 2010. lme4: Mixed-effects modeling with R. Http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf.

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

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

Biemann, Chris, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski & Torsten Zesch. 2013. Scalable con-

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

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

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.

Dąbrowska, Ewa. 2014. Words that go together: measuring individual differences in native speakers' knowledge of collocations. *The Mental Lexicon* 9(3). 401–418.

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

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. 2016a. 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, Ewa Dąbrowska & Antti Arppe. 2016b. Machine meets man: evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1). 1–33.

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.

Durrant, Philip & Alice Doherty. 2010. Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory* 6(2). 125–155.

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.

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

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

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.

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. More (old and new) misunderstandings of collostructional analysis: on Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536.

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

Gries, Stefan Th. 2015c. 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., Beate Hampe & Doris Schönefeld. 2005. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676.

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.

Gries, Stefan Th. & Stefanie Wulff. 2005. Do foreign language learners also have constructions? evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3. 182–200.

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

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.

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.

Johnson, Steven G. 2017. The nlopt nonlinear-optimization package. http://ab-initio.mit.edu/nlopt.

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

Kapatsinski, Vsevolod. 2014. What is grammar like? A usage-based constructionist perspective. *Linguistic Issues in Language Technology* 11. 1–41.

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).

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

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

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.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen & Douglas Bates. 2017. Balancing type I error and power in linear mixed models. *Journal of Memory and Language* 94. 305–315.

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

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

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

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

Mollin, Sandra. 2009. Combining corpus linguistic and psychological data on word co-occurrences: corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5(2). 175–200.

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

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

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.

Newman, John. 2011. Corpora and cognitive linguistics. *Revista Brasileira de Linguística Aplicada* 11(2). 521–559.

Newman, John & Tamara Sorenson Duncan. 2015. Convergence and divergence in cognitive linguistics: Facing up to alternative realities of linguistic catgeories. Talk given at the 13th international cognitive linguistics conference (ICLC-13).

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

R Core Team. 2014. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria.

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.

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

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

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. 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.

Stefanowitsch, Anatol & Susanne Flach. 2016. A corpus-based perspective on entrenchment. In Hans-Jörg Schmid (ed.), *Entrenchment, memory and automaticity. The psychology of linguistic knowledge and language learning*, 101–128. Berlin: De Gruyter.

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

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.

Taylor, John. 2008. Prototypes in cognitive linguistics. In Peter Robinson & Nick C. Ellis (eds.), *Handbook of cognitive linguistics and second language acquisition*, 39–65. New York and London: Routledge.

Taylor, John R. 2003. *Linguistic categorization*. Oxford: Oxford University Press 3rd edn.

Taylor, John R. 2012. *The mental corpus: how language is represented in the mind*. Oxford: Oxford University Press.

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.

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

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

Vasishth, Shravan. 2015. *A meta-analysis of relative clause processing in Mandarin Chinese using bias modelling*: School of Mathematics and Statistics of the University of Sheffield dissertation. http://www.ling.uni-potsdam.de/~vasishth/pdfs/VasishthMScStatistics.pdf.

Verbeemen, Timothy, Wolf Vanpaemel, Sven Pattyn, Gert Storms & Tom Verguts. 2007. Beyond exemplars and prototypes as memory representations of natural concepts: a clustering approach. *Journal of Memory and Language* 56(4). 537–554.

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

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

Wood, S. N. 2011. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 73(1). 3–36.

Zifonun, Gisela, Ludger Hoffmann & Bruno Strecker. 1997. *Grammatik der deutschen Sprache*, vol. 3. Berlin: De Gruyter.

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

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.