COGL.2017.0050

*Abstractions and exemplars: the measure noun phrase alternation in German*

formerly: *Competing Constructions for German measure NPs*

**Author’s replies to comments**

**Overview of the major changes**

I would like to thank all three reviewers and Dr Divjak for the highly helpful comments. I tried to implement as many of the suggestions as possible. Consequently, passages were removed, new ones added, and others moved to a different section and/or heavily edited. Also, the sectioning had to be changed significantly because of the addition of the corpus pre-study and the proposed clarifications in Section 2. For the convenience of the editors and possibly the reviewers (should they be asked to review the paper again), I begin by providing a list of the major changes, which are all based on feedback by the reviewers unless stated otherwise:

1. The title was changed to reflect the slightly changed focus of the paper.
2. The abstract was completely rewritten to reflect the slightly changed focus of the paper.
3. Section 1 was rewritten and significantly extended. This was due to the almost unanimous request for a stronger grounding in cognitive linguistics. I decided to focus on the prototype vs. exemplar debate and the issue of convergence between corpus-based and experimental findings.
4. The descriptive overview of the alternation is now in Section 2.2. It begins with a description of the alternation.
5. Section 2.1.2 about the syntax of the constructions was moved to its current place from the part where I develop the core hypotheses of the study (now in Section 2.2) because it really belongs more into the descriptive part in the context of this study.
6. In Section 2.1.3, I removed the discussion of the *von* and the the *voll(er)* constructions in order to make room for the additions recommended by the reviewers, and because these constructions are not really central to my point.
7. Section 2.2. now comes with two subsections, one about prototype effects (rewritten), and one about exemplar effects and multilevel modelling (newly written).
8. Section 2.2.1 was restructured and partly rewritten to provide a clearer description of the prototypes. The predicted usage profiles are now derived from the definition of the prototypes, not vice versa.
9. Section 2.2.2 was added because the exemplar perspective is an aspect that was neglected in the submitted version.
10. The old Section 3.3 was removed (Bayesian vs. maximum likelihood estimators) for space reasons, and to avoid unnecessary distractions from the core point of the paper.
11. Section 3.1.3 was extended to include some discussion of issues raised by the reviewers (operationalisation of the stylistic effects; collostructional attraction strength). It also had to be moved because it pertains to the new Section 3.2 (see immediately below) and Section 3.3.
12. Section 3.2 is based on a new auxiliary corpus study and was written from scratch in order to remedy shortcomings pointed out by the reviewers.
13. Given the increased amount of material reported in the revised paper, the sectioning in the part on the corpus studies was streamlined (Sampling and annotation; Statistical model; Interpretation) in order to improve the reading experience. The same goes for the two subsections of Section 4 (Setup, stimuli, and participants; Statistical model; Interpretation).
14. Section 3.3.2 contains a new paragraph where the results of using collexeme strength instead of the simple attraction quotient are reported.
15. Section 3.3.2 now closes with an interpretation of the differences between the (new) pre-study and the main study.
16. The introductory part of Section 4.1 now contains a discussion of why only a global validation of the corpus-based models was attempted.
17. Section 4.1.3 was mostly rewritten in order to provide a more differentiated discussion of the quality of the fit between corpus and experimental results. There is also a short discussion of why adding random slopes would lead to an overparametrised model not supported by the data.
18. **[This change is not based on reviewer feedback.]** Section 4.2.2 contains a new short paragraph where I report the results of fitting a GAMM to the reading data. I had done this anyway, and it fits in nicely with the second new major point in the revised paper, namely the quality of fit between corpus and experimental data – given that Divjak et al. (2016a) found that an analysis with a GAMM improved the fit for their reading time experiment.
19. Section 5 was completely rewritten to reflect the slightly changed focus of the paper.
20. **[This change is not based on reviewer feedback.]** I discovered a minor misspecification in the original GLMM for the corpus study, which was corrected. This did *not* change any of the results in a way which would affect the interpretation. However, the intercept estimate changed slightly, which offset the model predictions slightly, leading to a subsequent drop in the marginal R² for the forced-choice experiment. It is now slightly below 0.2, which was taken up in the discussion.

**Note: For all references used in these replies, the full citation**

**can be found in the revised version of my paper.**

COGL.2017.0050

*Abstractions and exemplars: the measure noun phrase alternation in German*

formerly: *Competing Constructions for German measure NPs*

**Author’s replies to comments**

**Replies to Dr Divjak’s comments**

For the convenience of the editors, I also reply to Dr Divjak's main points, although most of this is dealt with in detail again later. I ask that these comments be treated confidentially.

1. the paper should be more firmly grounded in cognitive (socio)linguistics, and the theory (PT or ET) should yield the hypothesis to test

I agree, and I found R2’s suggestion most plausible. She or he suggested that there is clear evidence in my data that something like a varying abstraction approach is more appropriate than a pure prototype approach. After all, I included both prototype *and* exemplar effects in my model anyway. Since this is an under-researched point (with rare exceptions like Divjak & Arppe 2013) of high importance in cognitive linguistics, I decided to make it one of the primary twists of the paper. Since my paper therefore stands in a direct line with other papers published in COGL (such as Divjak & Arppe 2013), I hope it is now sufficiently attractive to its readers.

Furthermore, the description of the prototypes was made much clearer.

Especially R3 – but also R2 – suggested that I should use constructs from Langacker’s Cognitive Grammar such as profiling or schemas. It is my impression that COGL is open to cognitive linguistics in the broad sense, and that it is not required to include CG modelling. Furthermore, I do not work in CG for various reasons, which is why I decided against such a modelling. However, the relation to CG is now made in the revised version (Section 1). There is also a brief comment in Section 2.2.1 (Footnote 13) which explains why I think a CG modelling would not help much in the present case. I kept this fairly backgrounded in a footnote because it is not one of my primary concerns.

2. the authors should be more explicit about the relation between the statistical model and the experimental findings - as pointed out, converging evidence in itself is no longer a new contribution to the field, and in this case the evidence might actually be diverging.

There is now extensive discussion of this matter both in the general Sections 1 and 5 and the report of the experiment (Section 4.1).

In general, however, I’d like to ask the editors to consider that in Divjak et al. (2016b, 3–4), Dr Divjak herself wrote that “only a *small* number of these corpus-based studies have been cross-validated” (highlighting mine; see full quote in l.1148–1153) in COGL. Thus, while cross-validation is not “new”, we are clearly nowhere near a conclusive picture, and a lot more studies are required.

Furthermore, not reaching convergence in a single study does in no way show that there is divergence, even with respect to the phenomenon and the experiment at hand. This is a basic fact of experimental quantitative work known at least since the work of Ronald A. Fisher in the 1920s and 1930s. This is why more studies and ultimately meta analyses are really needed. See also the comments in Section 5, especially Footnote 39.

That said, I agree wholeheartedly that the point was not adequately discussed in the submitted version at all, and the abovementioned amendments based on the reviewers' and Dr Divjak's comments have in my view helped to improve the paper a lot such that it now actually makes a contribution to the debate.

3. the authors should review the operationalization of their variables – reviewers #1 and #3 in particular provide detailed comments on the linguistic analysis

See my replies to the reviewers' comments below.

4. some aspects of the statistical analysis should be reconsidered – all choices (of frameworks, models and specifications) should be carefully justified; reviewer #2 provides a number of helpful suggestions.

See my replies to the R2's comments below. I also have some confidential remarks:

I removed the discussion of Bayesian inference completely, also because the paper is now a tad long anyway. However, I think that COGL misses a chance to raise the awareness of what different methods of statistical inference really mean and do (in a convenient way, by example). Simply following the Bayesian hype will not do the field any good in my opinion.

I have now included a short statement that my approach to statistical inference is Fisherian (Footnote 25 on p. 27), and I will not use Bayesian inference for the type of statistics presented here (as R2 suggests). Since COGL has no declared policy on statistical inference, I hope that this is in line with editorial policies.

R2 also suggests that I should use Bayesian estimators to solve convergence problems with random slopes in Sec. 4. I have now included a discussion of this in Section 4.1.3. In essence, I made some changes to achieve convergence using an ML estimator, which now converges but still *fails* because (predictably) the variance-covariance matrix is not estimated properly. Since Bates et al. (2015a) clearly state that this is a problem of model overparametrisation and data sparseness and *not* a problem of using an estimator which does better data dredging, I think using a Bayesian estimator would be ill-advised in this case (see Sec. 4.1.3). It must be stressed that among the authors of Bates et al. (2015a), we find the statistician who programmed the *lme4* package (Doug Bates) as well as two renowned linguists and statisticians (Shravan Vasishth and Harald Baayen).

5. Reviewers #1 and #3 are representative of the readership of our journal and suggest a number of changes to the text that will make it more accessible to our readers.

Please see my replies to the reviewers' comments below. I implemented the vast majority of the suggestions.

**Finally, if the text needs shortening, I suggest that the following passages can be removed because I am not sure that they are really relevant for most readers of COGL:**

1. Footnote 7, as I assume readers know about the difference between partitives and pseudo-partitives
2. Footnote 9
3. Footnote 12
4. Footnote 22
5. Footnote 23
6. Footnote 27
7. the part about why I used global model evaluation (“Readers might wonder [...]at least since the seminal Bresnan et al. (2007) paper.”) on p. 35
8. the part about random slopes (“In principle [...]to estimate the parameters of the more complex model with varying slopes”) on p. 39–40

COGL.2017.0050

*Abstractions and exemplars: the measure noun phrase alternation in German*

formerly: *Competing Constructions for German measure NPs*

**Author’s replies to comments**

**Replies to R1’s comments**

However, the introduction to and description of the phenomenon at issue, two (alternating) constructions for the expression of a particular partitive relationship in German, do not yet exhibit the clarity and systematic development expected. That means that revision is required for sections 1 and 2.

In particular, it is suggested that

- you re-arrange 2.1, so that your alternation is the starting point from which to elaborate anything relevant to its analysis; and

Section 2.1.1 now begins with a description of the alternation.

- re-structure 2.2 by giving the review of existing analyses in ONE part. Assessing them, formulate your own hypotheses and elaborate them accordingly.

I tried to implement this in the following way: I moved the part about the syntax to an extra subsection in the descriptive section (see the new Section 2.1.2). The part where my hypotheses are developed is now in Sections 2.2.1 and 2.2.2, separating the prototype effects and the exemplar effects. This is also a result of R2 and R3's comments. The internal structure of the text in 2.1.1 has been changed such that it begins with a clear statement about the prototypes and their meaning, with existing analyses providing the motivation in the remainder of the section.

I wonder if section 3.3 (the engaged argument about frequentist and Bayesian statistics) is needed for the study. The debate does not seem to be inherently related to the phenomenon under investigation, though it addresses a more general methodological issue, which certainly is not unimportant. Perhaps it could be 'backgrounded' and the convergence of the results of the MLE and MCMC methods be shown as a by-product of your linguistic analyses.

I removed the discussion for space reasons, and in order to comply with R2's comments.

Instead, the reader not familiar with such regression models will certainly benefit from a little more information on the models employed. It is suggested that, for example, some explanation be given in section 3.4 on what the values of the coefficients (in Table 3) mean.

I suspect it is inevitable that readers not familiar with generalised linear modelling will find some parts a rough read, and the length of the paper sadly forbids any further elaboration. However, I added comments about the advantages of multilevel models for the specific type of modelling in my paper in Section 2.2.2.

As regards this table, its caption contains very complex information that should be given (and explained) in the text.

I moved most of the information into the text body.

Moreover, it is also confusing that two columns have the same label (level). It is not obvious (to me) what the left-most ‘level’ is, the second ‘level’ seems to be the factor levels. This should be mended/clarified.

I corrected the table header, and with the elaboration in Section 2.2.2, I hope it is now clear which levels are referred to.

More generally, the (place of the) plots in section 3.4 is not always optimal, eg Fig 4 is mentioned on p 23, the figure is on p 25.

I tried to remedy this to the degree that the number of floats allowed it.

It is not always immediately obvious that the values given on the y-axes are scaled differently (it is pointed out only in Fig 2).

See Footnote 29 now.

In most plots, the print size (labels of the axes, for example) is very/too small. Additionally, what is specified in the headlines of the figures should go into their captions, part of the information in the captions should be taken up in the text (and be briefly explained). The abbreviations used NACadj and PGCadj are not used consistently, they need to be adapted in Table 2 and Fig 5 and 6. Fig 6 shows an interaction effect rather than ‘the main effect’.

All of this has now been fixed.

It is suggested that the text be proofread by a native speaker of English.

The submitted version had been read by a native speaker of Canadian English who is also a linguist. The new version has been proofread by the same person (except some parts in Sections 4 and 5 which I had to change again slightly after proofreading). If there are any specific parts which should be checked again, I would appreciate some hints as to where the remaining typos and grammar or style problems are.

In the print version available to me, object language is not marked (against metalanguage) - make sure that all example phrases/sentences are given in italics (or marked otherwise).

This might be a problem with online conversions going on in Manuscript Central or with a specific viewer. All meta language *is* in italics in the submitted PDF. Also, in the official PDF which I downloaded from Manuscript Central, everything looked fine as well.

The German examples are 'glossed' using ordinary English - which is problematic, eg for reasons of lost cases in English. How about using the Leipzig glosses for rendering the German expressions as precisely as possible?

I did not follow this recommendation because it would really have cluttered the examples. Also, I checked recent publications about German in COGL such as Madlener et al. (2017) in 28 (4), 757–798, and they do not use Leipzig glosses either. The relevant case is always marked in my examples at the level of phrases in the form of labelled brackets.

In the text, the word prototypical seems to be overused. Typical would be enough in many contexts, especially when talking about degrees: much more/very prototypical(ly) sounds odd

Yes, I removed a lot of *proto* in *prototypical*.

Make sure that the terms and abbreviations used in the figures and tables are given in a legend – or are explained in close vicinity.

In the revised version, I do not see any remaining abbreviations that need further explanation.

**The following list of minor suggestions was implemented, or the corresponding passage is no longer part of the paper as a result of the rewrite. I only comment on the points which I could not take into account:**

p 7: ... because the alternative cxns are not used in the same range of contexts --> which suggests that they are other, 'quasi-alternatives'?

I would not say that they are alternatives in a true sense. They are part of a loosely defined group of constructions expressing part-of and measurement relations, but – as I explained in the submitted version, they mean very different things. The *von* construction requires a determiner and is the dominant true partitive construction (notice that it also allows the core partitive use as in *drei von den Äpfeln* "three of the apples"), and the *voller* construction is only used with container nouns and always denotes the container, never the substance or the portion.

p. 8: ein Glass voll/voller roten Weins sounds fine with me; actually, the combination with container Ns seems to follow from the occurrence of voll/voller.

I just made an ad-hoc corpus search in the new DECOW16A, which has a lot of additional morphological annotation compared to DECOW14A, allowing users to make queries like

[word="voller"] [tag="ADJA"] [tag="NN" & morph="masc" & morph="gen" & morph="sg"]

The queries I made show the following picture: (1) the genitive occurs very productively with plural nouns, to a lesser degree with feminine singular nouns, and virtually never with masculine singular nouns. Also, it is mostly used predicatively and not attributively (*Die Schale war voller kleiner Äpfel.*). I agree, however, that the construction is much more acceptable than I made it sound in the submitted version.

In part due to these problems, I removed the paragraphs about the *voller* and the *von* constructions. Furthermore, it seemed to me to be the most irrelevant part for the actual point of the paper, and I needed to make room for other necessary elaborations based on the reviewers' comments.

p. 10: The part-of meanings of true partitives as in a slice of the cake represents the first stage of a development ... In German, the PGC is clearly the older construction (Zimmer, 2015) --> isn't that contradictory?

I would think that it makes sense the way I put it: The partitive comes first diachronically according to Koptjevskaja-Tamm. The PGC is older, and can still be used to express partitives while the NAC cannot.

p 16: Since the PGCadj contains a genitive itself, the regressor variable Genitive and the document-level variable Genitives are not fully independent. --> Is the document-level variable Genitives what you report as ‚badness‘ elsewhere?

The two are completely separate. The (unchanged) explanation is now found in Section 3.1.3.

COGL.2017.0050

*Abstractions and exemplars: the measure noun phrase alternation in German*

formerly: *Competing Constructions for German measure NPs*

**Author’s replies to comments**

**Replies to R2’s comments**

First, the interpretation in terms of Prototype Theory is unsatisfactory because of the following reasons:

a) The motivation for the use of that theory is wrong. The author writes on p. 4, “Prototype Theory (...) is preferred here simply because it fits the established alternation modeling paradigm”. I strongly believe that the causal link should be in the other direction. First, one chooses a theory to test, and then one selects an appropriate method. To choose a theory because it meets one’s R expertise is scientifically wrong.

I fully agree with the reviewer that the theoretical underpinnings of the paper needed clarification, and I made many changes to remedy this. I also thank her or him for pointing out that there was a highly relevant point (exemplars vs. prototypes) which I did not exploit in the submitted version.

However, I also want to point out that this was never about my R expertise. Prototype theory “fits the established alternation modelling paradigm” because the models are usually based on high-level abstract linguistic features. Rarely used exceptions are analogical models as popularised by Royal Skousen or recent approaches in Baayen et al. (2016) and Ramscar et al. (2016). Such approaches are *quite* rarely used in alternation research concerned with morpho-syntax, semantics, and pragmatics.

b) The author seems to speak about the prototypes of two superordinate constructions, which are labelled as NAC and PGC, e.g. p. 10, “In the remainder of this section, I argue that the NAC and the PGC are prototypes associated with different degrees of the grammaticalisation of the measure noun and related morpho-syntactic properties, as well as register effects. The degree of similarity of a given instance to either of the two prototypes makes speakers chose the NACadj or the PGCadj.”

However, the data include only a very restricted subset of Nmeasure + Nkind constructions: - only instances Nmeasure + Adj + Nkind

- only singular

- Nmeasure in genitive are excluded

- only with 100 most frequent mass nouns

While I understand the author’s reasons for using these data, he or she does not present empirical evidence that the distinctive features that are found in the subschemata with adjectives are also those that are distinctive of the superordinate constructions (i.e. the prototypes). We do not know if the features’ weights in the superordinate constructions are identical to or different from the weights in the subschemata with adjectives.

I fully agree that this was missing. In the revised version, Section 3.2 presents an new study of the non-alternating cases with mostly converging results. A comparison/discussion of the feature strengths is found towards the end of Section 3.3.3.

c). The author includes a predictor that clearly represents exemplar effects in the model. I have no problems with that, but it is strange that this obvious fact is ignored in the discussion. These are the relative frequencies with which measure noun lemmas and kind noun lemmas appear in the prototypical (non-alternating) PGCdet and NACbare. I also think that the effect of certain measure nouns on the choice of the constructions is difficult to interpret as prototype effects. By including random effects associated with the specific nouns, the author also models exemplar effects. This should be made clear.

I am grateful for this comment because it truly helped to make the paper more attractive. I rewrote Section 1 to include a proper discussion. Consequently, the title needed to be changed as well. Section 2.2 and Section 5 were also changed appropriately.

The second problem is the lack of novelty in a correlation between corpus-based results and experimental data. After numerous studies that displayed such convergence, it would be more interesting to focus on divergence and try to explain it (e.g. Newman & Sorenson Duncan 2015). Since the correlation is rather weak (judging from Rm in two models, which was about 0.2-0.3), one would need a deeper discussion of the discrepancies and their possible cognitive explanations.

I could not agree more. This was made the second important point of the paper. See Section 1, Section 4.1.3, Section 4.2.3, and Section 5. While some of the weakness of the R² coefficients could be traced back to a single badly chosen stimulus, I also think that we cannot expect a much higher R² if the conditional R² in the corpus study is lower than 0.5 to begin with. This point is also discussed in the paper.

It would also be interesting to test the effect of the predictors from the corpus- based model on the subjects’ responses directly, instead of using the predictions from the model trained on the corpus data. That would allow one to pinpoint the similarities and dissimilarities between the two methods.

I could not follow this suggestion for several reasons.

First, my statistical philosophy is Fisherian (see the new Footnote 25). Thus, I firmly believe that the way an experiment was designed predetermines the sample space, and I consequently cannot choose a method of statistical analysis freely *after* the experiment was conducted.

Second, the number of stimuli and the sample size are clearly much too small to squeeze anything useful about single regressors out of the data. This is now explained in the introduction to section 4.1.

Third, the global evaluation of corpus-derived models has a tradition in probabilistic grammar and cognitive linguistics, see also Section 4.1 (introductory part). The experiment was designed as part of this tradition.

Third, I find that the operationalization of some factors is not very careful. For example, the register is represented by two variables: ‘Badness’ and the normalized frequency of genitives in the corpus components. There seems to be no supporting evidence that these two can serve as reliable representatives of the register.

I would like to point out that the best part of the popular normative discussion about language in Germany is about the genitive (see Bastian Sick's literally best-selling pamphlet *Der Dativ ist dem Genitiv sein Tod* (part 1 and 2) and the linguistically informed reply by André Meinunger *Sick of Sick*), and that using the genitive is considered good style to such a degree that certain prepositions are now being used excessively with the genitive although they historically (and according to normative grammars) take a dative.

The Badness score is empirically evaluated against raters' judgements in Schäfer et al. (2013), which is referenced in the paper. I added a sentence about this in Section 3.1.3.

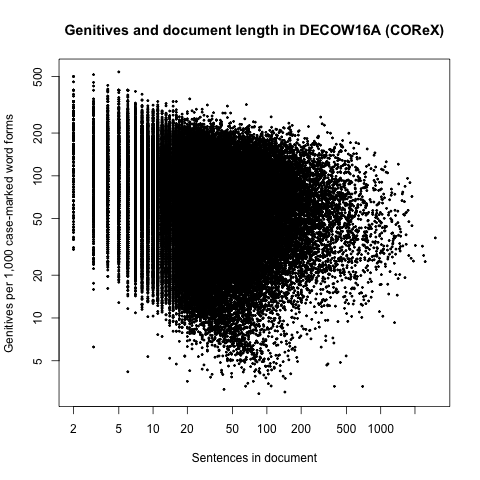
I changed Section 3.1.3 slightly, but I do not see any severe problems with the two proxy variables.

I agree, however, that *register* is probably the wrong term, and I changed it to *style*. This is mostly because using the term *register* (in my understanding) would require that the *social meaning* in the sense of Labov be specified, and I find this quite difficult. Style seems the much more neutral and modest term.

If the documents that constitute the web corpus are small (and I strongly suspect that they are), the

I cannot share this suspicion. Many documents in the corpus are quite long, for example because most forum and blog CMS serve archive versions of full discussion threads. In the beta version of the DECOW16 COReX meta data base (https://www.webcorpora.org/opendata/docdata/cow16a/), the mean document length is 816.5 words. (This is *words*, not tokens; also, it is the *usable* text, i.e., excluding boilerplate and noisy material. The COW creators have invested a lot of time and effort into creating cleaner corpora, and dubious material is no longer counted as good sentences.) In the long tail of the distribution, there are many very long documents, up to a maximal value of 46,739 words (longer than the target sample size for books in the BNC, which was 40,000).

frequency of genitives per document would not be a sufficiently sensitive measure because genitives are infrequent in present-day German.

The genitive is still the primary attributive case, and it is used with many prepositions. It reliably occurs with stable frequencies in any text which remotely resembles the written national standards (with a significantly lower frequency only in Swiss German because the native Alemannic dialect does not have a genitive anymore). Verb-dependent genitives have never been frequent, and they are virtually extinct. See Fleischer & Schallert (2011). See also Footnote 23 in the revised paper.

On the right, I include a scatter plot of the distribution of document lengths in sentences plotted against the genitive count per 1,000 case-marked forms in a sample of 82,000 documents from the COReX beta database (see Footnote 23 and immediately below). The distribution looks fine to me, maybe except for a minor overestimation of the genitive count in very short documents below ten sentences.

Notice that the document length measured in sentences is an ideal measure of reference here because usually (definitely in my corpus studies) complete sentences are queried. Documents with few sentences thus have a low chance of contributing anything to a concordance, and thus the slight bias in their *Genitives* score does not affect most corpus studies severely.

Finally, it is my impression that R1 and R3 are German and/or linguists working on German, and they did not object against my use of the *Genitives* score.

An alternative way would be to use Biber’s Multidimensional Analysis features, e.g. 1st and 2nd person pronouns, discourse markers, or average word length.

First of all, the genitive is *not* pathologically rare, as I pointed out above. Many documents, however, categorically do not contain first and second person pronouns. I do not think that using them would lead to better results.

Second, there is no corpus of German with Biber-style annotations available, and the proxy variables provided with DECOW14A such as the *Genitives* and *Badness* scores actually represented a huge step forward when they were introduced. At the time, there wasn't even a large German corpus with the required basic annotations (like morphological tagging) which could feed something like a Biber tagger. In the meantime, the COW creators spent two years adding rich basic annotations and creating something similar to the Biber tagger (the open-source COReX tagger) in collaboration with the Institut für Deutsche Sprache (IDS), Mannheim. Clearly, an effort like this cannot be expected from authors of research papers, especially since the full documents contained in the corpus cannot even be made available to them for download under EU intellectual property legislation. (There is no Free Use doctrine in Europe.) Therefore, as long as the COW project does not release a final version of the COReX database, nobody can user Biber-style features for any large corpus of German. The database is, however, not available in its final version as of today, and the COW creators advise against using it in published research, clearly warning that this would be bad scientific practice.

In sum, although I agree that the proxy variables are not optimal, they are the best available proxies. I would also like to point out that the *Genitives* variable in particular showed a considerable influence with narrow confidence intervals in the pre-study and the main study. This also contributes to substantiating its usefulness.

Moreover, it is not clear what is considered the high and low register – formality, more careful writing, or something else.

I agree this was unclear. I now call it a *stylistic* effect, see Sections 2.2.1 and 3.1.3.

Another issue is the use of frequencies of nouns in the superordinate constructions. There are many different measures that represent the entrenchment of collocates. The author rejects explicitly collostructional strength (Footnote 14), but unfortunately for the wrong reasons. First, he or she says that the goal here is “to quantify how often lemmas occur in the PGCdet and the NACbare, and these constructions do not compete at all but are rather mutually exclusive. Collostructional approaches are not made for such scenarios.” This statement is not true. One can use collostructional strength to measure the attraction between the construction and its slot fillers without any competition between two constructions (e.g. Stefanowitsch & Gries 2003). Second, the author claims that “the values resulting from collostructional analysis, i. e., logarithmised Fisher p-values, have a very unfavourable distribution in the case at hand. They cluster around 0 and they include values of −∞”. The problem of infinity values can be easily solved in R (e.g. package Rmpfr), and one usually gets reasonable scores.

While it was only mentioned in a footnote before, I now report the results of substituting the attraction variable with collexeme strength. It was a failure, and I discuss this in connection with the debate about convergence and non-convergence, which is now a prominent point in my paper thanks to the suggestions made by this reviewer. See Sections 3.1.3 and 3.3.2 and also Section 5.

Fourth, I don’t see any added value in comparing the MCMC and Maximum Likelihood models. Any introductory textbook will tell you that these models would yield very similar results if one uses weakly informative or flat priors. Nor does Levshina (2015), who the author refers to, claim that the results should be different, or that the Bayesian model should be more ‘precise’. The advantages are of the epistemological nature: most importantly, one can compute the posterior probabilities of the alternative hypothesis, instead of trying to reject the null hypothesis. As for what constitutes ‘true’ Bayesian statistics, whether one uses informative priors or not, there are different approaches in the literature, including linguistics. The use of non-informative or weakly informative priors seems to be the default approach in the broader Stan community, but their models do not cease to be Bayesian because of that.

The comparison was removed, and I will discuss these issues extensively elsewhere.

Fifth, the specification and diagnostics of the corpus-based model(s) should be described more precisely. Although I really liked the fact that the author used a bootstrap validation, I haven’t found a discussion of possible interactions between the predictors. Have they been tested?

I follow an approach which avoids model complexity for the sake of interpretability, and which only considers variables which have a theoretical motivation. It was clearly required in the associate editor's comments that “the theory should yield the hypothesis to test”. Thus, I am fundamentally opposed to upward model selection not guided by theoretical assumptions. See Footnote 28.

What about random slopes? The author also reports a convergence issue in Footnote 25. Has he or she tried to resolve this issue by trying different optimizers? Ironically, this is also a situation when Bayesian (or MCMC) modelling, which is rejected by the author, might come very handy! Moreover, I have a question regarding the evaluation of the model on p.30.

It was my impression that it is not standard practice in COGL to report details about the optimiser used. While I always use an alternative BOBYQA from the *nloptr* package with lme4 estimates for corpus studies (see the new Footnote 24), I had not done so for the models in Section 4. I re-estimated both models using this BOBYQA implementation, which in my experience is the fastest and most robust optimiser – and almost a requirement if one uses the bootstrap because all other optimisers are too slow.

This gave me the chance to mention Bates et al. (2015a) and Matuschek et al. (2017), who argue (contra Barr et al., 2013) that overparametrised mixed models with random slopes are often not supported by the data. In my case, the covariance parameters were estimated at -1 and 1, which is a clear sign of this. Bates et al. (2015a) also argue that this is a problem of the parametrisation of the model given the data, and that it cannot be solved by using a different estimator like MCMC. See Section 4.1.3.

The author says it has high quality. What are the criteria? Judging from the low pseudo-R values, one can conclude that some important factors are missing.

The formulation was obviously too optimistic, and I toned it down.

COGL.2017.0050

*Abstractions and exemplars: the measure noun phrase alternation in German*

formerly: *Competing Constructions for German measure NPs*

**Author’s replies to comments**

**Replies to R3’s comments**

1. Other factors favouring one of the options

I suggest to include (or at least mention) another factor favouring NACadj over PGCadj. In the case of highly lexicalised adj-noun combinations in the kind-noun component (e.g. saure Gurken, weiße Bohnen, schwarzer Tee, rote Bete), there is, as far as I can see, a more than strong preference for NACadj, to a degree that PGC is often odd or receives a quite different interpretation.

(1) a. Er kaufte ein Glas saure Gurken.

Er bestellte einen Teller weiße Bohnen.

b. ??? Er kaufte ein Glas saurer Gurken.

??? Er bestellte einen Teller weißer Bohnen.

In the study, the author shows the somehow parallel preference for strongly grammaticalised measure nouns (e.g. Pfund, Gramm, Meter) to enter NACadj (one of the factors tested). I would add the above factor ‘lexicalisation of the adj/n combination’ for consideration as it nicely complements (taking somewhat the inverse perspective on the construction) the case of the factor ‘grammaticalisation of the measure noun’. Since lexicalised adj-noun combinations are cognitively speaking more rigid units (entrenched) than non-lexicalised ones, the NAC construction is closer to their status (the status of saure Gurken / weiße Bohnen being close to the one of genuine compounds such as Rotwein, Weißwein) and hence preferred. When placed in the PGC, these items are not quite interpreted in terms of their entrenched lexicalised meaning. Hence, this observation supports the overall argument made by the author and also provides elements for a description of the schematic meaning of these constructions (see comment 3 below).

The author includes one sociolinguistic factor in his/her analysis, i.e. register. It seems to me that another sociolinguistic factor, i.e. regional variation, would need to be considered or at least mentioned. This would add to the usage-based commitment of the study. I am unaware of corpus data on this issue, but I do have the impression that the preference for one of the alternatives (NACadj / PGCadj) differs from region to region in addition to a high-low register profile.

I agree with the reviewer that both factors might influence the alternation. I discuss the first one (lexicalised A+N combinations) in Section 5, and I mention the second one (regional variation) in Footnote 14.

The lexicalisation of A+N combinations was not taken into account in the corpus study in the revised version because it would have meant several weeks of work with uncertain outcome. (After all, we only have an intuition to begin with.) Also, while it would be an interesting effect, I am not sure whether it would fit into the picture of the prototype painted in the paper, and the effect could be orthogonal to the prototype effects. But I will definitely work on this, probably in an experimental paradigm.

2. Narrow Apposition Construction (NAC)

The author gives the following characterisation of the syntactic conditions in NACbare: ”If the kind noun is bare – i.e., if it comes neither with a determiner nor a modifying adjective – it has to agree in case with the measure noun” (p.4 line 27). I hesitate to agree with this characterisation. Instantiations of this construction with verbs requiring Dative or Genitive suggest that it is not a “typical” narrow apposition with case identity:

(2) a. Wir bedürfen eines Glases Wein.

b. \*Wir bedürfen eines Glases Weins.

Perhaps, (2b) can be ruled out by a general “quirky constraint” mentioned by the author “that genitive NPs in German require the presence of some strongly case-marked element” (p.5 line 16). However, (3b) with an archaic Dative inflection (-e) added to avoid the otherwise indistinguishable case situation sounds very odd. Dative Plural, which still has a fairly distinct inflection in German, is more than odd (3c). (3d) suggests a similar (though not identical) situation for instances where the NP receives case (Dative) from a preposition.

(3) a. Ich vertraue dem Glas Wein.

b. ??? Ich vertraue der Flasche Weine.

c. \*Ich folge dem Sack Äpfeln.

d. ?? mit einer Flasche Weine / einem Sack Äpfeln

I.e., I fail to see strict case identity in this construction. If one extends the “quirky” constraint on the Genitive (see above) to the Dative in order to accommodate the case-identity claim or adds further constraints (e.g. phonological ones or constraints linked to grammatical gender), the entire notion of case identity is somewhat void. Eisenberg (1989: 258) argues instead that the historical case forms (being barely distinct or not distinct at all) were reanalysed as a Nominative (or alternatively, the kind noun in this construction is regarded as caseless/uninflected). The interpretation in terms of a Nominative can also be found in other grammars (e.g. Jung 1982: §253). Zifunun, Hoffmann & Strecker (Vol. 3, 1997: 1980) prefer the “uninflected nomen-invarians” analysis. I am not saying that e.g. the Nominative interpretation is correct (I think it is problematic); however, as noted above, I fail to see strong case identity.

In the NACadj (4a/b), however, there is a rather strong tendency to case identity indeed.

(4) a. Ich vertraue dem Glas rotem Wein(e) / der Flasche rotem Wein(e).

b. mit einem Glas rotem Wein(e) / einer Flasche rotem Wein(e)

Zifunun, Hoffmann & Strecker (Vol. 3, 1997: 1981) also assume case identity here. Eisenberg (1989: 259f.) takes a similar stance; however, he also points to some more or less acceptable alternatives and argues that the grammatical status of the kind noun (case identity?, Nominative?, Accusative?) is not fixed/clear in contemporary German. I agree that he (partly) overstates this point (esp. for the Nominative). However, I think that the case of NACadj with highly lexicalised adj-noun combinations in the kind-noun component (cf. comment 1) is again interesting in this respect. All of the options in (5) are acceptable to me (and also to other native speakers of German that I consulted) and get a fair amount of hits in a quick unsystematic google search (perhaps mit einer Tasse schwarzer Tee (NOM) is also acceptable to some speakers; it is not quite attested by a google search):

(5) a. mit einer Tasse schwarzem Tee (DAT)

b. mit einer Tasse schwarzen Tee (AKK)

c. mit einer Tasse schwarzen Tees (GEN, PGC); very few tokens

This is not meant to add confusion to the picture. Rather, it is the behaviour one would expect if the strict case-identity criterion for NACbare is given up (see above). Strongly lexicalised adj-noun combinations in the kind-noun component hesitate between a pattern closer to the NACbare prototype (weak case identity) and NACadj (rather strong case identity).

In sum, I see no need for the author to insist on the case-identity criterion for NACbare, neither in the light of the data nor in the light of his/her modelling. It is clearly compatible with his/her argument to assume that NACadj is linked to the NACbare prototype and largely preserves/requires original case identity, while case identity in turn has been (largely) given up in NACbare (if it ever had been strong in the first place). This, however, would need to be shown diachronically (e.g. in a follow-up study).

I fully agree. The situation is really more complicated than I made it sound. In Section 2.1.1, I now discuss this in some detail, including additional evidence from weak nouns which I came across and found very confusing myself. This is also something I would be curious to do more research on because my own intuitions are quite hazy.

Furthermore, there are interesting effects on preference/acceptability depending on the item from which the measure noun receives case (a verb vs. a preposition). From a short list of examples with Dative verbs and Dative prepositions I fail to see any straightforward pattern. Clearly, a corpus study is promising in this respect (e.g. in a follow-up investigation).

This is another convincing idea. However, I think such an effect would surely be independent of the prototype and exemplar effects on which I focus, much like the *Measurecase* variable. I therefore added this to my list of aspects to explore in the future.

3. (Schematic) meaning(s) of the constructions

The “meaning” of the two alternating constructions is spelled out by the author (i) relative to the usage profiles of the NACbare and PGCdet prototypes and (ii) in terms of their own usage profiles. Section 2 provides a descriptive account of their formal properties. All of that is convincing (except for the postulated case identity in NACbare; see comment 2 above). What I miss, however, is a fully spelled-out proposal as to a cognitive-linguistic modelling of the distinct schematic meaning(s) of these two constructions beyond their usage patterns. The author states that “subtle” semantic differences exist between them. Can/should this difference be reduced to different usage profiles (favouring factors)? How can these differences be described beyond an account of usage profiles? I.e., eventually, do these constructions have a distinct schematic meaning or can they be captured by their usage profiles alone? The reader expects an account of the schematic meaning of the constructions since the author explicitly favours prototype theory (assuming higher-level schematic meaning) over exemplar-based models (that try to do without such abstractions).

The paper contains several potential starting points for such a “positive” modelling of the schematic meaning(s) of these constructions. For instance, the author observes that NAC “forms the more prototypical environment for expressing measurement (as opposed to partitivity)” (page 10, line 36), that PGC encodes a “part-of relation” while NAC is "more closely associated with the quantity-of relation" (p. 10, 11). I suggest that the paper is slightly extended to provide a fully spelled-out CL modelling of these observations, e.g. in terms of different profiling (in the sense of Langacker), i.e. in terms of the relative weight of the components. Eisenberg’s (1989: section 7.3.2) analysis in terms of different dependency relations (yielding different types of appositional relations) could provide further elements of such a modelling. The comments made above on the behaviour of strongly lexicalised adj-noun combinations in the kind-noun component are also worth considering in this context.

Again, I found this comment to be very helpful. I do not work in Cognitive Grammar in the narrow sense, and I also found it difficult to formulate a schematic meaning because none of the constructions is exclusively associated with a unique meaning (which is problematic given the discreteness of schema membership). However, I now define the prototype clearly including an account of the relation between the alternating NACadj and PGCadj and the non-alternating NACbare and PGCdet. The expected usage profiles are derived from the prototypical meanings. See Section 2.2.1 and especially Footnote 13.

MINOR ISSUES

All of these were rectified.