Prototypes in the univerbation of German verb-noun units

Roland Schäfer
Deutsche Sprache und Linguistik,
Humboldt-Universität zu Berlin
Dorotheenstraße 24, 10117 Berlin
schaefer@hu-berlin.de

Ulrike Sayatz
Deutsche und niederl. Philologie,
Freie Universität Berlin
Habelschwerdter Allee 45, 14195 Berlin
sayatz@fu-berlin.de

Abstract …

Keywords: univerbation, prototypes, production experiments, corpus data, German

1 The form and history of noun-verb units in German

2 The status of noun-verb units?

3 Corpus-based analysis of the usage of verb-noun units

3.1 Design and choice of corpus

The goal of the corpus study was to assess (i) which V + N units exist in written German usage, and (ii) how strongly they are attracted by the univerbation effect. The operationalisation of question (ii) relied on the fact that the major graphemic principles in German are clear and dominant, and that they are both deeply rooted in diachrony and well entrenched in writers' usage. The relevant major principle for the present study was compound spelling of words, which we took as an indication that in the grammars of the writer the compounded words had single-word status.

Research questions (i) and (ii) – as opposed to the – are clearly not driven by strong hypotheses derived from theory, and we consequently adopted a

data-driven approach with a post-hoc interpretation of the results. [1] Hence, we needed to extract (close to) *all* relevant N + V units from an ideally very large and varied corpus as a first step. In a second step, we had to count their occurrences in compound and separate spelling in the relevant morphosyntactic contexts enumerated in Section 1, viz. as the heads of noun phrases, *am* progressives, as participles in analytical verb forms, and as infinitives in a range of verbal constructions (for example with modal verbs).

Clearly, a large corpus with rich morphological and morphosyntactic annotations containing texts written in a broad variety of registers and styles (including ones with low normative pressure) was required. We chose the DECOW16B corpus (Schäfer & Bildhauer 2012) because it fulfils all the abovementioned criteria.[2] Much like the SketchEngine corpora (Kilgarriff et al. 2014), the COW corpora contain web documents from recent years. However, the German DECOW (containing 20.5 billion tokens in 808 million sentences and 17.1 million documents) offers a much wider range of annotations compared to SketchEngine corpora, including morphological annotations and serveal levels of syntactic annotation (dependencies and topological parses). For our purpose, the fully internal analysis of nominal compounds described in Schäfer & Pankratz (2018) was particularly of interest. It allows for searches of roots within nominal compounds. For example, we could query compounds with a deverbal head such as *Zeitnehmen* ('time taking'). Furthermore, the interface offered by the creators of the COW corpora allows for automated queries controlled by Python scripts using the open-source *SeaCOW* interface.[3] The scripts we used to make the queries are released on a cureted open-data server along with all data as well as the LaTeX, knitr, and R scripts created in the writing of this paper.[4]

3.2 Sampling and annotation

The first step of the implementation of the corpus study was the generation of a list of actually occurring N + V units. We obtained such a list by querying for compounds with a nominal non-head and a deverbal head. (See the scripts available under the abovementioned DOI for concrete queties and further details.) The rationale behind this approach was that any N + V unit of interest should occur at least one in compound spelling as a fully

[1] The results obtained from the corpus were also used in the choice of the stimuly for the experiment reported in Section 4.
[2] https://www.webcorpora.org
[3] https://github.com/rsling/seacow
[4] The DOI of the data set will be revealed in the accepted version of this paper.

nominalised compound. Since this step relied on automatic annotation, the results contained erroneous results, which we cleaned through manual annotation. The resulting list contained 820 N + V units.

In the second step, we created lists of all relevant inflectional forms of the verb in each V + N unit and used these to query all possible compound and separate spellings (including variance in capitalisation) of each of the 820 N + V units. In total, 28,700 queries were executed to create the final data set used here, a number which clearly demonstrates the necessity of script-based corpus access in data-driven methods. The queries were matched by 1,029,190 compound spellings and 1,292,886 separate spellings, which results in a total sample size of 2,322,076.[5]

For each N + V unit in the sample, the following variables were annotated automatically: (i) the verb, (ii) the noun, (iii) whether a linking element is used in the use as a full noun, (iv) the overall frequency in the corpus. Additionally, we manually coded all 820 N + V units for the relation holding between the verb and the noun (see Section 1). The codes used in clearcut cases were *Object* (442 units) and *Adjunct* (286 units). For 92 units, both relations were conceivable, and those cases were coded as *Undetermined*. This class is illustrated by *Daumenlutschen* ("thumb sucking"), which could be paraphrased as either (1a) or (1b).

(1) a. das Lutschen des Daumens
the sucking of the thumb

b. das Lutschen am Daumen
the sucking on the thumb

3.3 Modelling the corpus data

In this section, we present the parameter estimates (and predictions of conditional modes) for a multilevel generalised model (or generalised linear mixed model, GLMM) which models the – in our view – the relevant factors influencing speakers' choice of the compound and the separate spelling.[6] Given the grand total of 2,322,076 observations in the sample (see Section **??**), we will completely refrain of using inferential statistics per se. For samples of such magnitude in data-driven approaches, frequentist significance tests are the wrong tool. Bayesian methods reliably converge with

[5] Notice that two highly frequent N + V units were excluded because they could be considered outliers, having an overly strong tendency to be used in compound spelling. They are *Teilnehmen* ("taking part") and *Maßnehmen* ("taking measure").

[6] See (Schäfer n.d.) for an overview of the method and our philosophy in modelling.

frequentist methods at this sample size. Therefore, we provide standard likelihood ratio confidence intervals for parameter estimates and prediction intervals for conditional modes as an approximate measure quantifying the precision of the parameter estimates. The models we specify reflect theoretically motivated decisions, and we therefore reject all types of model selection by means of step-up or step-down procedures.

As argued in Section 2, we expect the probability of the univerbation of N + V units to depend on the morphosyntactic context, the relation holding between the verb and the noun, the presence of absence of a linking element in the nominal compound (as a marker of a stronger reconceptualisation) and on the specific N + V unit (a lexical tendency). Accordingly, the response variable was chosen to be the proportion of compound spellings among all spellings of the N + V unit. The input data frame to the estimator was thus a table of 820 proportions, one for each N + V unit.[7] We specified four regressors. The only first-level (or observation-level) fixed effect regressor is the morphosyntactic context (a four-way categorical variable). As there is a huge number of 820 N + V units, the lexical indicator variable for the individual N + V unit should not be used as a fixed effect (Gelman & Hill 2006: 244–247). Therefore, we specified a generalised linear model with the N + V unit variable as a random effect. The variables encoding the internal relation and the presence/absence of a linking element are nested inside the levels of the random effect, and they are therefore threated as second-level fixed effects in a multilevel model.In R notation, the specification is shown in Equation 2.[8]

(2) $\texttt{Proportion} \sim \texttt{Context} + \texttt{Relation} + \texttt{Link} + (1|\texttt{NVUnit})$

The estimated parameters of the model are given in Table 1. Additionally, effect plots for *Context* and *Relation* are given in Figure 11.[9] As expected, the prototypically verbal contexts (infinitives and participles in ana-

[7] Binomial models can be specified in this manner (Zuur et al. 2009: 245–260). In the estimation of such models, the influence of each proportion is weighted according to the number of cases observed to calculate it. Without the weighting, highly frequent observed proportions would have too little influence on the estimation of the model, and infrequent ones would have an inappropriately high influence. In the case at hand, such a model on proportion data is also a convenient way of getting around difficulties of estimating a model on the raw 2,322,076 observations.

[8] See Appendix A for a precise specification in mathematical notation.

[9] Put in an oversimplified manner, effect plots for binomial GLM(M)s (Fox & Weisberg 2018) plot the probability of the outcome across values of a regressor assuming default values for all other regressors. While model coefficients in binomial (and other) models have no direct

	Estimate	CI low	CI high
(Intercept)	-3.584	-3.584	-3.584
ContextParticiple	-0.084	-0.084	-0.084
ContextNP	2.682	2.682	2.682
ContextProgressive	3.714	3.714	3.714
RelationUndetermined	1.332	1.332	1.332
RelationAdjunct	3.110	3.110	3.110
LinkYes	0.354	0.354	0.354

Table 1: Coefficient table for the binomial GLMM modelling the corpus data with 95% profile likelihood ratio confidence intervals. The horizontal line separates first-level and second-level effects. Weighting was used to account for the bias in models on proportion data. Random effect for V+N lemma: Intercept $= 4.425$, sd $= 2.103$. The intercepts model the fixed effects Relation=Object and Link=No. Nakagawa & Schielzeth's $R^2_m = 0.519$ and $R^2_c = 0.999$.

lytic verb forms) are associated with a low probability of compound spelling (the infinitive is on the intercept -3.584, and participles have a coefficient of -0.084). NPs and progressives as prototypically nominal contexts clearly favour compound spelling (coefficients of 2.682 and 3.714, respectively). Both the coefficients and the effect plot (right panel in Figure 1) show alow probability of compound spelling when the relation between the verb and the noun (on the intercept) is that of an object, and a high probability when the relation is that of an adjunct (coefficient 3.110). The undetermined cases are in between the two clearcut cases (coefficient 1.332). The presence of a linking element in fully nominalised compounds favours compound spelling only slightly (coefficient 0.354).

Given the narrow confidence intervals and the high marginal measure of determination $R^2_m = 0.519$, we consider the hypotheses regarding fixed effects as well corroborated by the data, especially the effects of the context and the internal relation. Based on our commitment to a usage-based probabilistic view of language, we also predicted differences between N + V units not explainable by the fixed effects. These effects would show up as the residual variance in the random effects (in the form of the conditional modes) not modelled by the second-level effects. The conditional modes are

interpretation in terms of probability, effect plots allow a more intuitive interpretation in terms of changes in probability.

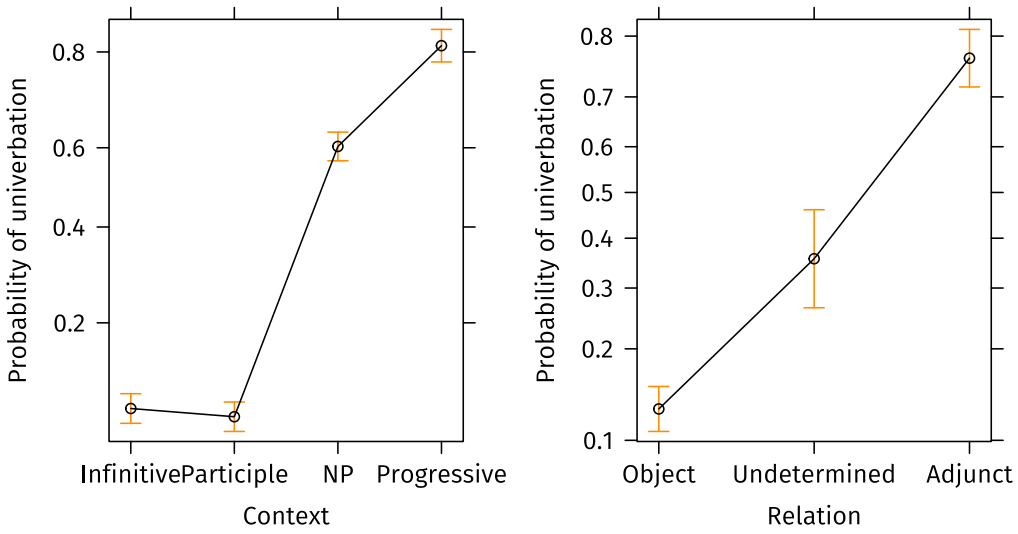

Figure 1: Effect plot for the regressor encoding the morphosyntactic context of the N+V unit and the regressor encoding the syntactic relation within the N+V unit in the GLMM modelling the corpus data.

centered around a second-level intercept of 4.425 with a standard deviation of 2.103. The standard deviation is a sign that there is considerable variation between single N + V units. Furthermore, the conditional is as high as $R^2_c = 0.999$. This is standardly interpreted as saying that the fixed effects and the idiosyncatic effect of concrete N + V units almost fully explain the variance in the data. A random selection of 20 conditional modes, which corroborates this interpretation through obvious differences with mostly very narrow prediction intervals, is shown in Figure 2.

The individual V + N unit thus plays a major role in writers' affinity to the univerbation of V + N units. This was shown in the form of the second-level predictors and the residual conditional modes. In Section 3.4, we approach this effect using yet another method, and the results obtained using that method will be used to predict participants' behaviour in the controlled experiment reported in Section 4.

3.4 Association strengths

In this section, we report an analysis of the item-specific affinities of N + V units towards univerbation. The reasons for this additional analysis of the

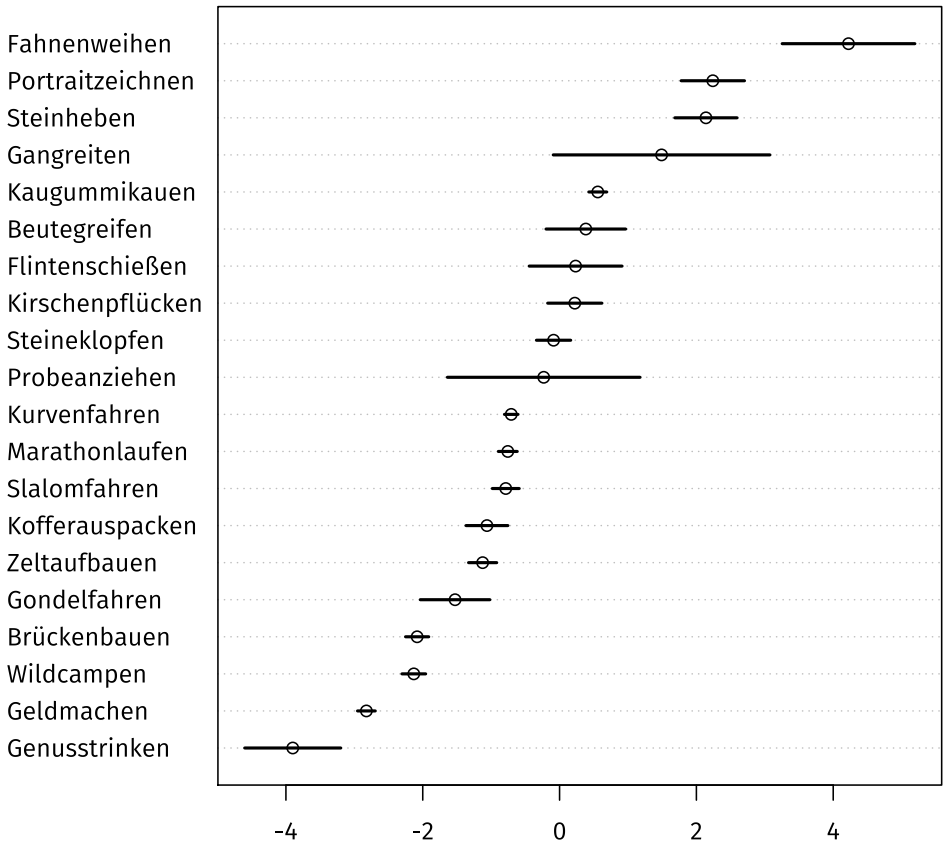

Figure 2: A random selection of conditional modes with 95% prediction intervals for the levels of the random effect in the GLMM modelling the corpus data.

	Univerbation	No univerbation
Specific N+V unit	c_{11}	c_{21}
All other N+V Units	c_{21}	c_{22}

Table 2: 2×2 contingency table as used in the calculation of the strengths of the associations of N+V units with univerbation.

data is twofold. First, we aim to demonstrate that the same interpretation can be obtained using a method that is technically much simpler and more robust against problems with the distribution of the data and against misinterpretation than multilevel modelling. This is a valuable contribution to the current discussisons in linguistics and statistics, also in the sense of methodological pluralism (see, for example Arppe & Järvikivi 2007). Second, we saw in Section 3.3 that the second-level predictors and the individual N+V units – both being related to the choice of concrete N+V units – are highly predictive of the outcome (univerbation or not). Therefore, in the experiment reported in Section 4, we need to control for the N+V units' affinity towards univerbation. The measures introduced here are ideally suited for this task.

The method we use seems superficially similar to collocational analysis (Evert 2008 for an overview) or collostructional analysis (Stefanowitsch & Gries 2003). However, there are major differences. We were interested in a quantification of how strongly single N+V units tended towards univerbation vis-a-vis all other N+V units. Thus, we need to compare the count of cases with univerbation of each N+V units versus the count of cases without univerbation with the same counts for all other N+V units. Such comparisons must be made relative to the overall number of the specific N+V units and all others, and the relevant counts are nicely summarised in a 2×2 contingency table shown in Table **??**.

We're interested in deviations between the first row and the second row, and there is a range of statistical measures for that. One can, for example, use odds ratios or effects strengths from frequentist statistical tests.[10] We chose Cramér's v derived from standard χ^2 scores ($v = \sqrt{\chi^2/n}$). The v measure quantifies how strongly the counts deviate from a situation where there is no difference between the individual N+V unit (cells c_{11} and c_{21})

[10] p-values from frequentist statistical tests are measures of evidende, and therefore not appropriate in such situations (Schmid & Küchenhoff 2013; Küchenhoff & Schmid 2015) although they were used in early collostructional analysis. However, even collostructional analysis is now mostly used with measures of effect strength (Gries 2015).

and all other N + V units (cells c_{21} and c_{22}). Since Cramér's v always is in the range between 0 and 1, it allows us to compare analyses where the sample size is different. In itself, v does not tell us whether the deviation is negative (for a N + V unit with less than average compound spellings) or positive (for a N + V unit with more than average compound spellings). The information about the direction of the deviation is added by multiplying v with the sign of the upper left cell of the residual table of the χ^2 test. The association scores are related to the second-level model (including the conditional modes), but they have a much more accessible interpretation.

We calculated the signed v for each of the 820 N + V units. Their distribution is plotted in the form of a density estimate in Figure 3.[11]

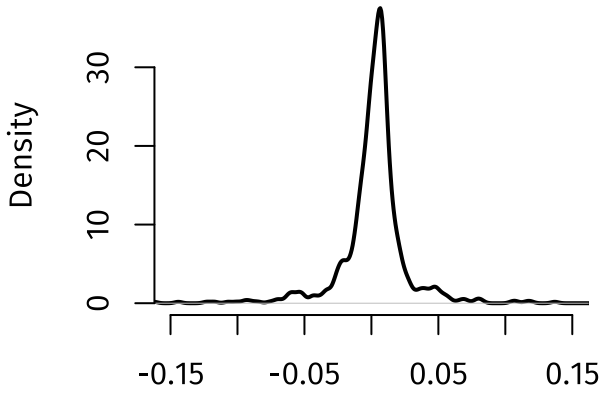

Figure 3: Density estimate of the distribution of the overall association scores (across all morphosyntactic conditions) with n = 820.

Based on the annotations in the corpus data set, we can also compare the association strengths for specific morphosyntactic contexts. The counts as shown in Table **??** are simply reduced to the counts in the four contexts. With the resulting lower sample sizes, the χ^2 measure can no longer be calculated in a number of cases, leading to lower n_{Unit}. The resulting distributions are shown in Figure 4.

[11] It approximates a scaled symmetric χ^2 distribution squashed between -1 and 1.

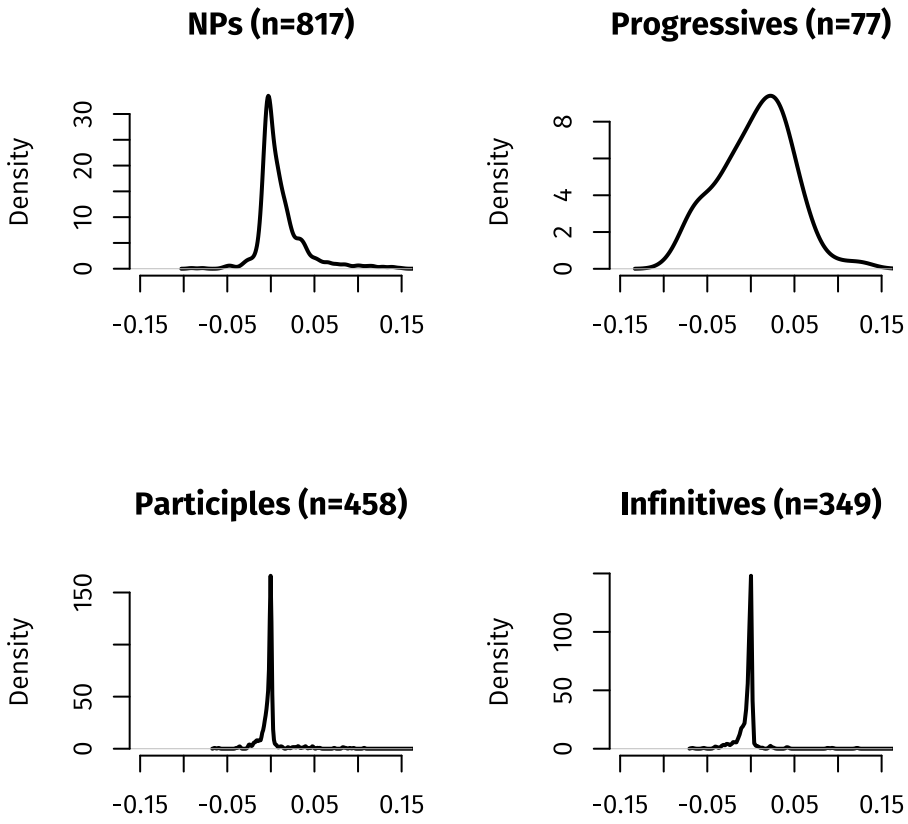

Figure 4: Density estimates of the distribution of the association scores in the specific morphosyntactic conditions..

The context-wise distributions of the association scores corroborate the results from the GLMM reported in Section 3.3. In the NP context (top left panel of Figure 4), the right tail of the curve is much heavier than the left tail, which means there are mostly higher than usual tendencies towards univerbation. In the syntactically similar progressive (top right panel), the distribution is (very) approximately symmetric, but given the low number of 77 N + V units for which v could be calculated, the result cannot be seen as stable.[12] Both prototypically verbal contexts (lower two panels) show heavy left tails, meaning that N + V units tend to resist univerbation in these contexts. Once again, this is just another (and maybe more intuitive) look at the data in addition the GLMM analysis.

For the selection of stimuli in the experiment, the overall association strength (Figure 3) is relevant, because it truely represents the effect of the unit, independently of the context. The context effect will be controlled independently in the experiment. To illustrate how the data anlysis allows for a selection of N + V units based on their affinity towards univerbation, we show the top ten units with the highest negative and highest positive association in Table **??**.

V+N Unit	Assoc.	Rel.	V+N Unit	Assoc.	Rel.
Teilhaben	0.18	Object	Gedankenmachen	-0.16	Object
Radfahren	0.18	N/D	Geldverdienen	-0.14	Object
Computerspielen	0.14	Adjunct	Rechtgeben	-0.12	Object
Zeitreisen	0.12	Adjunct	Spaßhaben	-0.12	Object
Skifahren	0.12	Adjunct	Rechthaben	-0.11	Object
Autofahren	0.11	N/D	Kinderhaben	-0.10	Object
Probefahren	0.11	Adjunct	Zeitnehmen	-0.09	Object
Bogenschießen	0.08	N/D	Auftraggeben	-0.09	Object
Schifffahren	0.08	N/D	Fehlermachen	-0.09	Object
Windsurfen	0.08	Adjunct	Urlaubmachen	-0.08	Object

Table 3: Top ten V+N units with a strong tendency for univerbation (left panel) and top ten V+N units with a strong tendency against univerbation (right panel).

[12] The low number is one the one hand due to the fact that progressives are rare compared to NPs, participles, and infinitives. On the other hand, it is likely that many N + V units cannot be used in the progressive for semantic or pragmatic reasons. The data set created by us would allow us to go into a detailed analysis of this question, but we postpone this for later due to space constraints.

The tables illustrate that units with the strongest tendencies against univerbation are predominantly ones with an object relation. The ones which most strongly favour univerbation are mostly ones with an adjunct relation or an ambiguous relation. The ten items with the least clear tendency in either direction are shown in Table **??**. They mostly have an internal object relation.

V+N Unit	Assoc.	Rel.
Klavierspielen	0.01	Object
Theaterspielen	0.01	N/D
Filmemachen	0.01	Object
Autowaschen	0.01	Object
Zigarettenrauchen	0.01	Object
Haarewaschen	0.00	Object
Notenlesen	0.00	Object
Golfspielen	-0.00	Object
Wasserholen	-0.01	Object
Haareschneiden	-0.01	Object

Table 4: Top ten V+N units without any tendency for or against univerbation.

Among the units with an object relation, it is difficult to tell based on native-speaker intuition, why the ones in Table **??** should have no preference and the ones in Table **??** should resist univerbation. While we can model the tendencies to a large extent using linguistic features, there are obvious item-specific effects which should be taken seriously from a theoretical perspective, and which must be accounted for in behavioural experiments. We now turn to such an experiment in Section 4.

4 Elicited production of noun-verb units in written language

5 Explaining the process of noun-verb univerbation

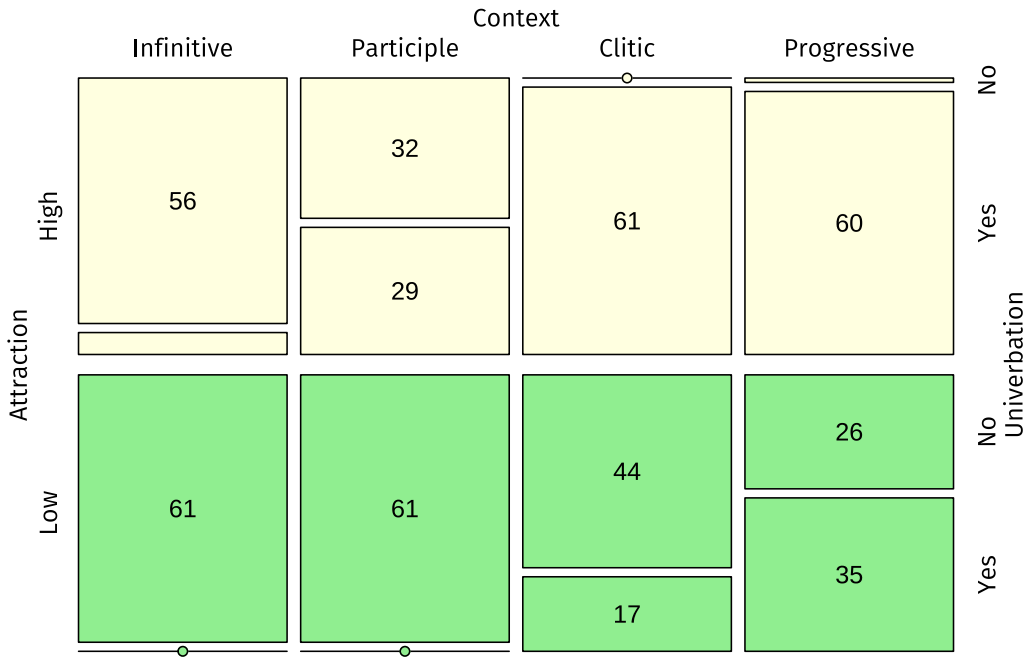

Figure 5: Mosaic plot of the responses in the production experiment (vertical right) grouped by the morphosyntactic context (horizontal) and the binned N+V unit's attraction strength calculated from the corpus (vertical left).

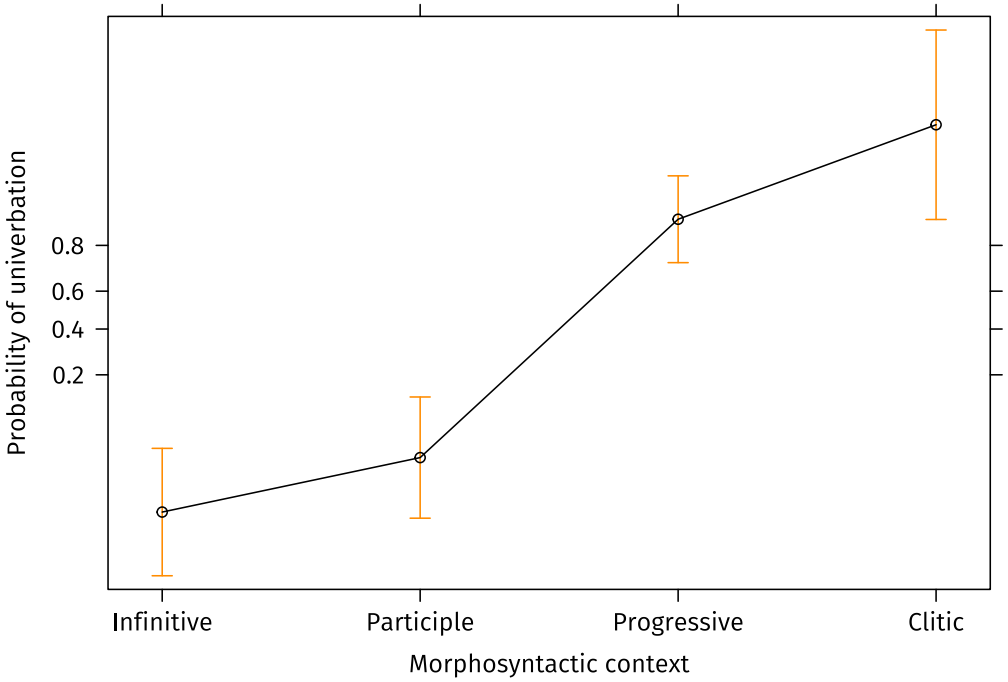

Figure 6: Effect plot for the regressor encoding the morphosyntactic context in the GLMM modelling the experimental data.

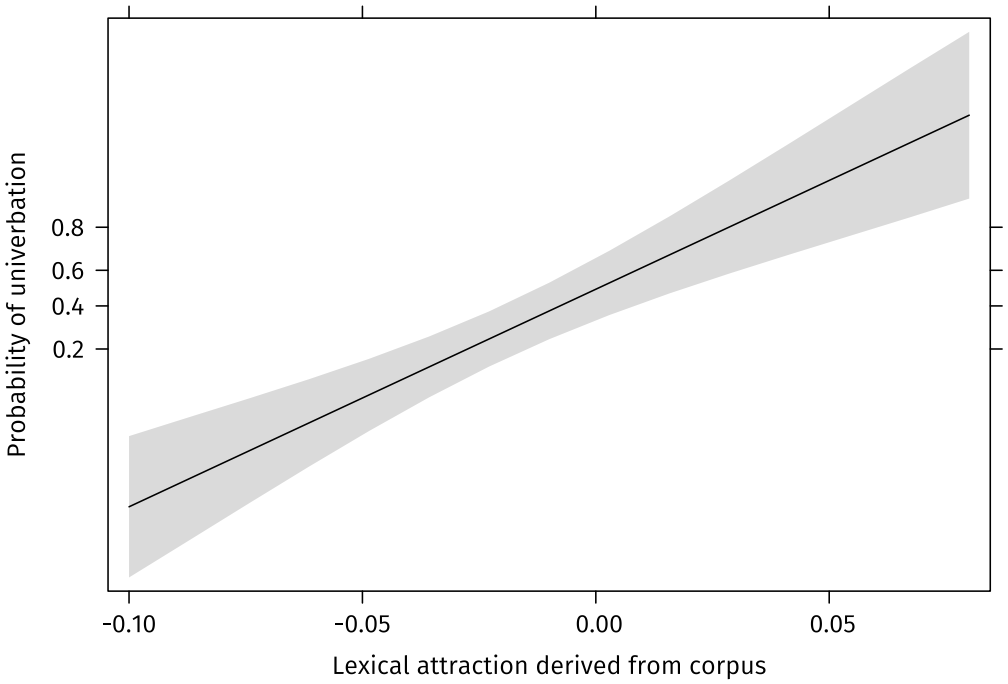

Figure 7: Effect plot for the regressor encoding the N+V unit's corpus-derived association with univerbation in the GLMM modelling the experimental data.

	Estimate	CI low	CI high
(Intercept)	-3.960	-5.460	-2.790
AttractionNum	49.541	35.193	74.789
ContextParticiple	1.167	-0.324	2.614
ContextProgressive	6.273	4.730	8.249
ContextClitic	8.297	6.071	11.720

Table 5: Coefficient table for the GLMM modelling the experiment data with 95% confidence intervals.Nakagawa & Schielzeth's $R^2_m = 0.804$ and $R^2_c = 0.897$. Random effect for participant: Intercept $= 2.967$, sd $= 1.723$.

Acknowledgments

A Full specification of the corpus GLMM

In Section 3.3, the specification of the model was gicen in Equation 2, repeated here as Equation 3.

$$\texttt{Proportion} \sim \texttt{Context} + \texttt{Relation} + \texttt{Link} + (1|\texttt{NV}) \tag{3}$$

This notation blurs the difference between first-level and second-level fixed effects. The model specification is the crucial step in statistical modelling since it encodes the researchers' commitment to a causal mechanism controlling the phenomenon to be modelled (in this case, writers' mental grammars with respect to the univerbation of N + V units). Model specification thus deserves more attention than 3 has to offer. Mathematically and thus more transparently, the model is given in Equation 4. The notation with angled brackets in $\alpha_{NV_j[i]}$ should be read as "the value of the random effect α_{NV} for the factor level j, chosen appropriately for observation i.

$$Prop_{Comp_i} = logit^{-1}[\alpha_0 + \alpha_{NV_j[i]} + \vec{\beta}_{Context} \cdot \vec{x}_{Context_i}] \tag{4}$$

The proportion of compound spellings $Prop_{Comp_i}$ is the logit-transformed sum of the overall intercept α_0, the random intercept for the j-th N + V unit $\alpha_{NV_j[i]}$ (whichever is observed in observation i) and the dot product of the vector of dummy-coded binary value for the morphosyntactic context $\vec{x}_{Context_i}$ and the vector of their corresponding regressors $\vec{\beta}_{Context}$. Since

it is a multilevel model, α_{NV} has its own linear model, which is given in Equation 5.

$$\alpha_{NV_j} = \gamma_j + \vec{\delta}_{Relation} \cdot \vec{x}_{Relation_j} + \delta_{Link} \cdot x_{Link_j} \tag{5}$$

It is assumed that Equation 6 holds.

$$\alpha_{NV} \sim Norm \tag{6}$$

The random effects are assumed to be a normally distributed variable α_{NV} which is for each N + V unit j given as the sum of the conditional mode of unit i (often wrongly called the *random effect* per se), the dot product $\vec{\delta}_{Relation} \cdot \vec{x}_{Relation_j}$ of the vector of binary variables encoding the relation and the vector of their corresponding coefficients and finally the product $\delta_{Link} \cdot x_{Link}$ of the binary variable encoding the presence of a linking element and its coefficient.

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. http://dx.doi.org/10.1515/cllt.2007.009.

Evert, Stefan. (2008). Corpora and collocations. In Anke Lüdeling & Maria Kytö (eds.), *Corpus linguistics. an international handbook*, 1212–1248. Berlin: Mouton. http://dx.doi.org/10.1515/9783110213881.2.1212.

Fox, John & Sanford Weisberg. (2018). Visualizing fit and lack of fit in complex regression models: effect plots with partial residuals. *Journal of Statistical Software* 87(9). 1–27.

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

Gries, Stefan Th. (2015). More (old and new) misunderstandings of collostructional analysis: on Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý & Vít Suchomel. (2014). The Sketch Engine: ten years on. *Lexicography*. 1–30. http://dx.doi.org/10.1007/s40607-014-0009-9.

Küchenhoff, Helmut & Hans-Jörg Schmid. (2015). Reply to “More (old and new) misunderstandings of collostructional analysis: on Schmid & Küchenhoff” by Stefan Th. Gries. *Cognitive Linguistics* 26(3). 537–547.

Schäfer, Roland. (N.d.). Statistische Inferenz in der Linguistik. in preparation.

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 eighth international conference on language resources and evaluation (LREC ’12) international conference on language resources and evaluation (LREC 12)*, 486–493. Istanbul: European Language Resources Association (ELRA).

Schäfer, Roland & Elizabeth Pankratz. (2018). The plural interpretability of German linking elements. *Morphology* 28(4). 325–358. http://dx.doi.org/10.1007/s11525-018-9331-5.

Schmid, Hans-Jörg & Helmut Küchenhoff. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577. http://dx.doi.org/10.1515/cog-2013-0018.

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

Zuur, Alain F., Elena N. Ieno, Neil Walker, Anatoly A. Saveliev & Graham M. Smith. (2009). *Mixed effects models and extensions in ecology with R*. Berlin etc.: Springer. http://dx.doi.org/10.1007/978-0-387-87458-6.