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 Analysing 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

[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

open-source SeaCOW interface.[3] The scripts we used to make the queries are released on a curated 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 nominalised compound. Since this step relied on automatic annotation, the results contained erroneous results, which we cleaned through manual annotation. The resulting list contained 819 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 819 N + V units. In total, 28,665 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 958,118 compound spellings and 1,288,768 separate spellings, which results in a total sample size of 2,246,886.[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 819 N + V units for the relation holding between the verb and the noun (see Section 1). The codes used in clearcut cases were *Object* (441 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

[3] https://github.com/rsling/seacow
[4] The DOI of the data set will be revealed in the accepted version of this paper.
[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”).

b. das Lutschen am Daumen
the sucking on the thumb

3.3 Results 1: Multilevel model

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 2,246,886 observations in the sample (see Section 3.2), 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 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 819 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 819 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

[6] See (Schäfer 2020, to appear) for an overview of the method and our philosophy in modelling.

[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 2,246,886 observations.

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 (2).[8]

(2) `Proportion` ∼ (1|`NVUnit`) + `Context` + `Relation` + `Link`

	Estimate	CI low	CI high
(Intercept)	-4.787	-4.787	-4.787
ContextParticiple	1.054	1.054	1.054
ContextNP	3.886	3.886	3.886
ContextProgressive	4.907	4.907	4.907
RelationUndetermined	1.339	1.339	1.339
RelationAdjunct	3.132	3.132	3.132
LinkYes	0.361	0.361	0.361

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 = 4.430, sd = 2.105. The intercepts model the fixed effects Relation=Object and Link=No. Nakagawa & Schielzeth's $R^2_m = 0.577$ and $R^2_c = 0.999$.

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 analytic verb forms) are associated with a low probability of compound spelling (the infinitive is on the intercept −4.787, and participles have a coefficient of 1.054). NPs and progressives as prototypically nominal contexts clearly favour compound spelling (coefficients of 3.886 and 4.907, respectively).

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

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 3.132). The undetermined cases are in between the two clearcut cases (coefficient 1.339). The presence of a linking element in fully nominalised compounds favours compound spelling only slightly (coefficient 0.361).

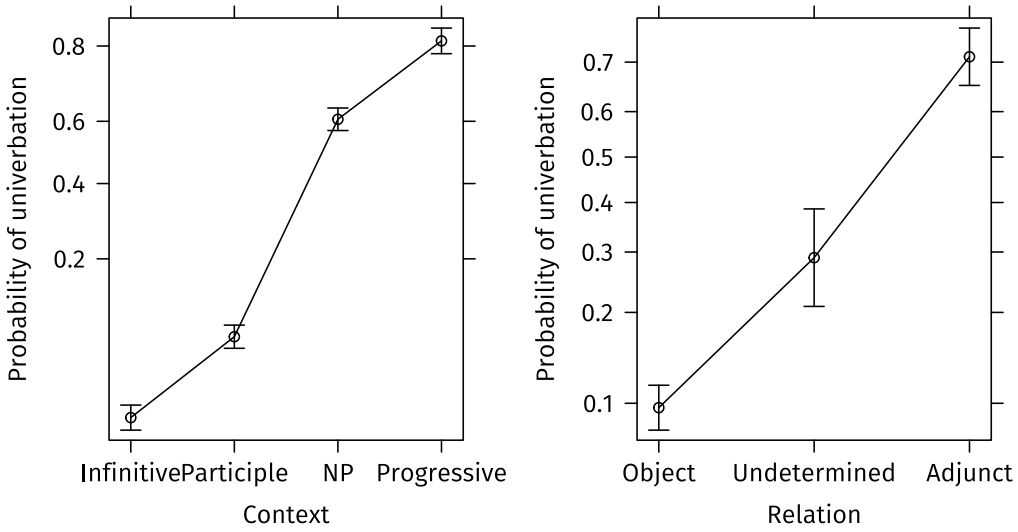

Figure 1: Effect plots for the regressor encoding the morphosyntactic context of the N+V unit and the regressor encoding the syntactic relation within the N+V unit in the GLMM modelling the corpus data.

Given the narrow confidence intervals and the high marginal measure of determination $R^2_m = 0.577$, 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 centered around a second-level intercept of 4.430 with a standard deviation of 2.105. 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.

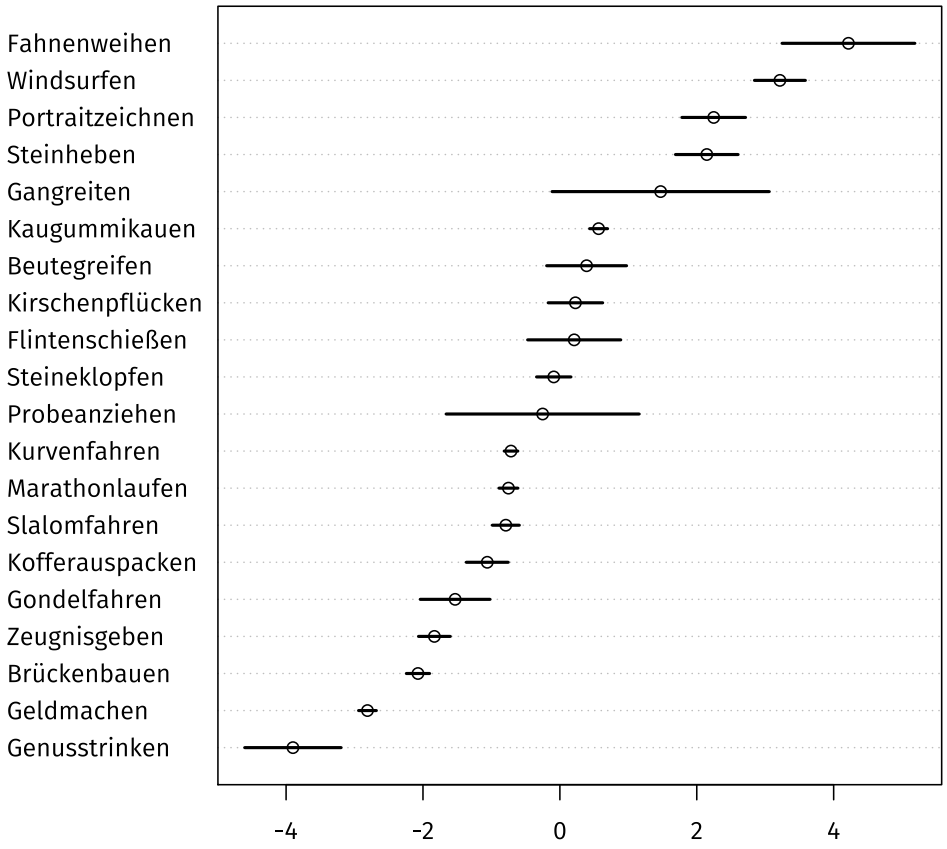

Figure 2: A random selection of conditional modes with 95% prediction intervals for the levels of the random effect in the GLMM modelling the corpus data.

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 Results 2: 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 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 2.

	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.

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}) 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.

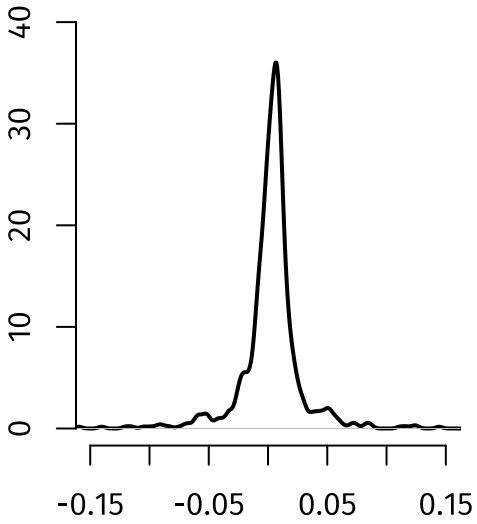

Figure 3: Density estimate of the distribution of the overall association scores (across all morphosyntactic conditions) with $n_{Unit} = 819$.

We calculated the signed v for each of the 819 N+V units. Their distribution is plotted in the form of a density estimate in Figure 3.[11] See also (Schäfer & Pankratz 2018) for similar use of association measures.

[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).

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

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 2 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.

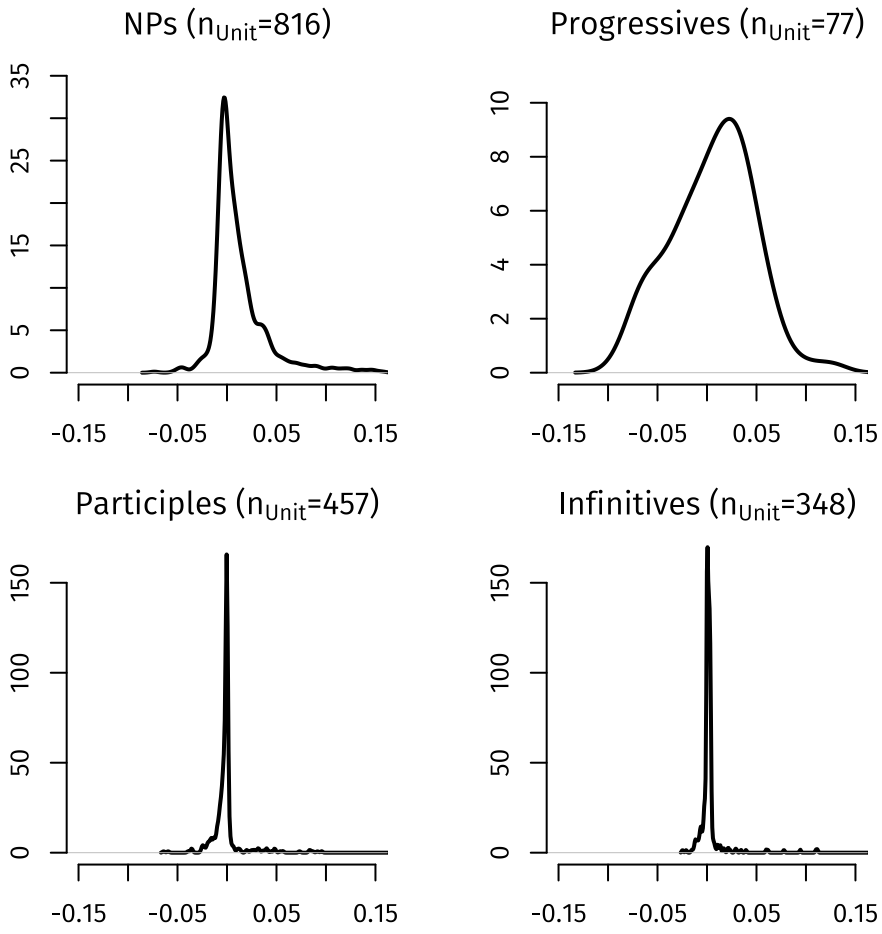

Figure 4: Density estimates of the distribution of the association scores in the specific morphosyntactic conditions..

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 ν 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 3.

V+N Unit	Assoc.	Rel.	V+N Unit	Assoc.	Rel.
Radfahren	0.190	N/D	Gedankenmachen	-0.160	Object
Computerspielen	0.144	Adjunct	Geldverdienen	-0.140	Object
Zeitreisen	0.125	Adjunct	Rechtgeben	-0.120	Object
Skifahren	0.123	Adjunct	Spaßhaben	-0.115	Object
Autofahren	0.117	N/D	Rechthaben	-0.105	Object
Probefahren	0.111	Adjunct	Kinderhaben	-0.099	Object
Bogenschießen	0.087	N/D	Zeitnehmen	-0.093	Object
Schifffahren	0.085	N/D	Auftraggeben	-0.092	Object
Windsurfen	0.084	Adjunct	Fehlermachen	-0.088	Object
Bergsteigen	0.082	Adjunct	Urlaubmachen	-0.083	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).

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 4. They mostly have an internal object relation.

[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.

V+N Unit	Assoc.	Rel.
Autowaschen	0.009	Object
Zigarettenrauchen	0.007	Object
Haarewaschen	0.005	Object
Notenlesen	0.003	Object
Golfspielen	0.001	Object
Haareschneiden	-0.007	Object
Wasserholen	-0.008	Object
Feuermachen	-0.009	Object
Blutabnehmen	-0.009	Object
Schlangestehen	-0.010	Adjunct

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 4 should have no preference and the ones in Table 3 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

4.1 Design and participants

The goal of the experiment was to corroborate the findings from the corpus and to test whether writers' behaviour under controlled experimental conditions is similar to the behaviour of writers under uncontrolled circumstances. Furthermore, the experiment allowed us to quantify individual preferences, something which is impossible with most types of corpus data.

We used pre-recorded auditory stimuli in the experiment in order to elicit spellings of given N + V units. The stimuli were chosen based on theoretical criteria and information obtained from the corpus. We constructed eight sentences instantiation four contexts, twice each. In each context, we chose one N + V unit with a high and one with a low attraction strength ac-

cording to the corpus.[13] The sentences were constructed in a ways such that all N + V units were the predicate of a subordinate clause in order to consistently ensure verb-last constituent order and avoid interferring verb-second effects, which are inevitable in independent sentences. The full sentences are given in Appendix B.

We added 32 fillers, resulting in a total of fourty sentences being read to the participants.[14] Of the fourty sentences, twenty (including the target items, of course) had to be written down. The order of the target items was randomised, by it was made sure that there were at least three sentences in between pairs of items. There were nine distractors in the form of yes-no questions related to random sentences previously heard by the participants. An overview of the item design is shown in Table 5, where each line represents the features of one of the eight items.

Context	N+V unit	Attraction
Infinitive	Platzmachen	-0.052
	Seilspringen	0.011
Participle	Mutmachen	-0.069
	Probehören	0.055
Progressive	Teetrinken	-0.037
	Bogenschießen	0.087
Clitic	Spaßhaben	-0.115
	Bergsteigen	0.082

Table 5: Items from the experiment, chosen by context and attraction score.

In total, 61 participants took part in the experiment, all of them first-semester students of German Language and Literature at Freie Universität Berlin. They were between 18 and 44 years old with a median age of 22 years. There were two separate groups (32 and 29 participants, respectively), and the randomisation of the stimuli was different between the two groups.

[13] Given the overall constraints on the choice of the items, *low* and *high* had to be seen as relative terms. However, all low attractions scores are higher than zero, all high attraction scores are greater than zero, and for each context, the pair of low and high attraction scores differs at least by 0.05.

[14] Of the fourty fillers, six were items from an unrelated experiment.

4.2 Results

The distribution of responses of the experiment is shown in the form of a mosaic plot in Figure 5. It shows the number of compound spellings (univerbation) and separate spellings in each of the four contexts and for N + V units with high and low attraction score.

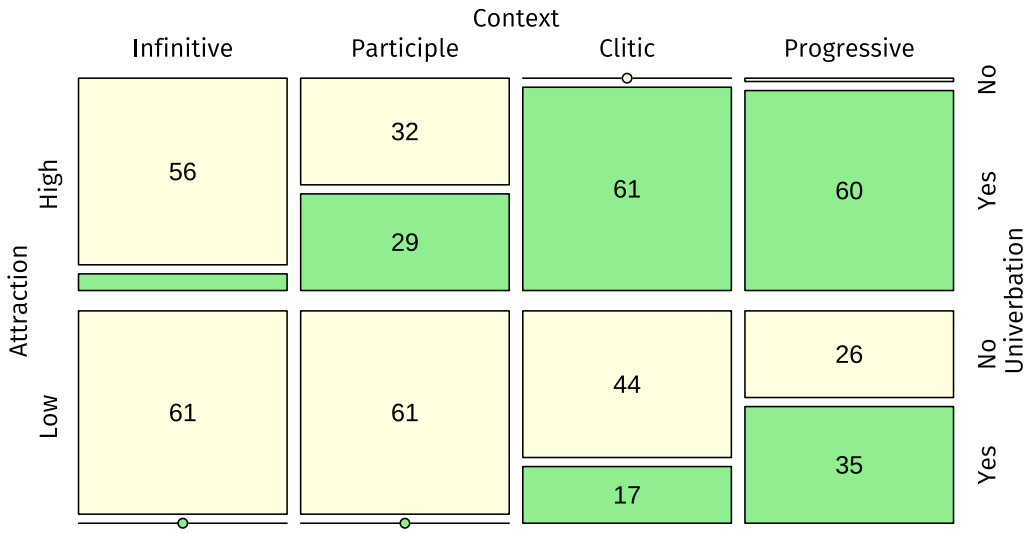

Figure 5: Mosaic plot of the responses in the production experiment (vertical right) grouped by the morphosyntactic context (horizontal) and the binned N+V unit's attraction strength calculated from the corpus (vertical left).

The overall number of positive responses (i. e., , compound spellings) rises across the four contexts. It's 5 for the infinitive, 29 for the participle, 78 for NPs with cliticised article, and finally 95 for the progressive (in each case out of 122. For the N + V units with a high attraction score, participants (almost) always use compound spelling in NPs with a cliticised article (61 out of 61) and in the progressive (60). Between the infinitive (5) and the participle (29), there is a clear differentiation in positive responses, however.

For the N + V units with a low attraction score, the items with an infinitive (0) or a participle (0) don't seem to allow univerbation at all. However, with NPs (17) and progressives (35), we see a considerable number of positive responses.

Clearly, both independent variables are highly useful in predicting the behaviour of participants. However, among the items with low association scores, we would expect the NPs as highly prototypical nominal contexts to trigger univerbation most strongly, while in the experiment they lose to the progressive (17 out of 61 for the NPs, 35 for the progressives). To examine the results further, we proceeded to estimate the parameters of a GLMM with additional control for individual participants in the form of a random effect. Instead of using a grouping variable for the N + V units, we included their association strengths directly in the model. The model specification in R notation is given in (3). Appendix C provides the specification mathematical notation. The coefficient estimates for the GLMM are reported in Table 6.

(3) `Univerbation` ∼ (1|`Participant`) + `Attraction` + `Context`

	Estimate	CI low	CI high
(Intercept)	-4.026	-5.534	-2.851
Attraction	48.740	34.657	73.476
ContextParticiple	1.126	-0.375	2.574
ContextProgressive	6.224	4.691	8.184
ContextClitic	8.166	5.978	11.513

Table 6: Coefficient table for the GLMM modelling the experiment data with 95% confidence intervals.Nakagawa & Schielzeth's $R^2_m = 0.803$ and $R^2_c = 0.896$. Random effect for participant: Intercept = 2.966, sd = 1.722 The intercept models the fixed effect Context=Infinitive as well as Attraction=0..

There is some variation between writers as captured in the standard deviation of the conditional modes (1.722), but the small difference between the very high marginal R^2 (0.803) and the conditional R^2 (0.896) suggests that speaker variation does not help to explain much of the variance in the the data. The coefficients indicate that the attraction strength derived from the corpus is positively correlated with participants tendency to univerbate (48.740). Given the large confidence interval ([−0.375..2.574]), there seems to be no evidence that the participle has a different effect than the infinitive (which is on the intercept). On the other hand, progressives (6.224) and NPs with cliticised articles (8.166) clearly have a much more positive effect on

the probability of univerbation. Thus, in the GLMM analysis NPs appear to have a stronger tendency to favour univerbation than progressives. This is in line with our theoretical predictions but seems to contradict the descriptive analysis (see Figure 5). For the further analysis of this apparent contradiction, we provide effect plots for the two fixed effects in Figure 6.

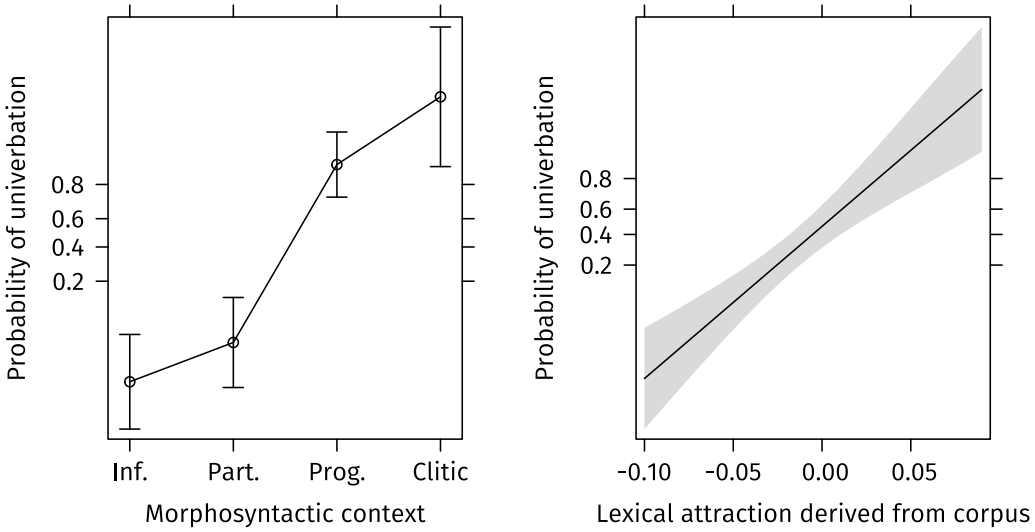

Figure 6: Effect plots for the regressor encoding the morphosyntactic context and the attraction strength as calculated from the corpus in the GLMM modelling the experimental data.

The effect plots show the same picture as the coefficient table. The prototypically verbal contexts are associated with low probabilities of univerbation, the two prototypically nominal ones with high probabilities of univerbation. While progressives and NPs with clitics show the order predicted by theory, there is only very weak evidence that the difference is substantial (large confidence intervals). The attraction scores are neatly correlated with the probability of univerbation.

The apparent paradox with respect to the order of the effects of NP and progressive contexts that we see between Figure 5 on the one hand and Table 6 and Figure 6 on the other hand can be explained by looking at the concrete attraction strengths in Table 5. The unit with "low" attraction used in the progressive context (*Teetrinken*) has a numeric attraction score of −0.037, which is much closer to 0 than the one used in the NP context (*Spaßhaben*) with −0.115. At the same time, the high attraction counter-

parts are rather close numerically (0.087 for *Bogenschießen* and 0.082 for *Bergsteigen*). Figure 5 therefore shows an item-specific positive bias for the progressive context, which is likely due to the concrete choice of items. The truely multifactorial analysis in the form of a GLMM compensates for it because it uses the numerical attraction scores. As the selection of stimuli is often not possible with perfect control over all variables, the more advanced statistical analysis protects us against misinterpretation.

In sum, the experiment clearly corroborates our theoretically motivated predictions and the corpus study. We proceed to a final analysis of the phenomenon at hand in Section 5.

5 Explaining noun-verb univerbation

Acknowledgments

We thank Luise Rissmann for her help annotating and cleaning the corpus data.

A Full specification of the corpus GLMM

In Section 3.3, the specification of the model was given in R notation as (2), repeated here as (4).

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

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 4 has to offer. Mathematically and thus more transparently, the model is given in (5). 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.

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

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 (6). It is also assumed that 7 holds.

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

(7) $$\alpha_{NV} \sim Norm$$

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.

B Sentences used in the experiment

The N + V units are typeset in smallcaps and spelled as separate words. The order of the sentences corresponds to Table 5.

(8) Lara trat zur Seite, um PLATZ zu MACHEN.
Lara stepped to the side in order room to make
Lara stepped aside to make way.

(9) Sarah ging auf den Spielplatz, um SEIL zu SPRINGEN.
Sarah went onto the playground in order rope to jump
Sarah went to the playground to do some skipping.

(10) Leon konnte nur deshalb gewinnen, weil Johanna ihm
Leon could only therefore win because Johanna him
MUT GEMACHT hat.
courage made has
Leon could win only because Johanna encouraged him.

(11) Maria hat einen Kopfhörer gekauft, nachdem sie ihn PROBE
Maria has a headphone bought after she it test
GEHÖRT hatte.
listened had
Maria bought a headphone after doing a listening test.

(12) Melanie mag Fußball, weil es ein Sport zum SPASS HABEN ist.
Melanie likes soccer because it a sport to the fun have is
Melanie likes soccer because it's a fun sport.

(13) Benjamin ruft seinen Freund an, weil er eine Frage zum
Banjamin calls his friend on because he a quaestion to the
BERG STEIGEN hat.
mountain climbing has
Benjamin calls his firend because he has a question about mountain climbing.

(14) Kim sah sich das Tennisspiel an, solange sie am TEE
Kim watched herself the tennis match on while she at the tea
TRINKEN war.
drink was
Kim watched the tennis match while drinking some tea.

(15) Simone hört ein Hörbuch, während sie am BOGEN
Simone listens an audiobook while she at the bow
SCHIESSEN ist.
shoot is
Simone listened to an audiobook while practicing archery.

C Full specification of the experiment GLMM

The model is specified in the same notation as in Appendix A in (16). The regressor $x_{Attract_i}$ is numeric (the attraction score), whereas $\vec{x}_{Context_i}$ is a dummy-coded vector of binary variables. Additionally, it is assumed that (17) holds.

(16) $$Pr(Univ = 1) = logit^{-1}[\alpha_0 + \alpha_{Part_j[i]} + \beta_{Attract} \cdot x_{Attract_i} + \vec{\beta}_{Context} \cdot \vec{x}_{Context_i}]$$

(17) $$\alpha_{Participant} \sim Norm$$

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. (2020, to appear). Generalized linear mixed models. In Stefan Gries & Magali Paquot (eds.), *The practical handbook of corpus linguistics*. Berlin, Heidelberg: Springer.

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.