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 simple standard frequentist 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 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 three regressors. 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.[8] The estimated parameters of the model are given in Table 1.

As expected,

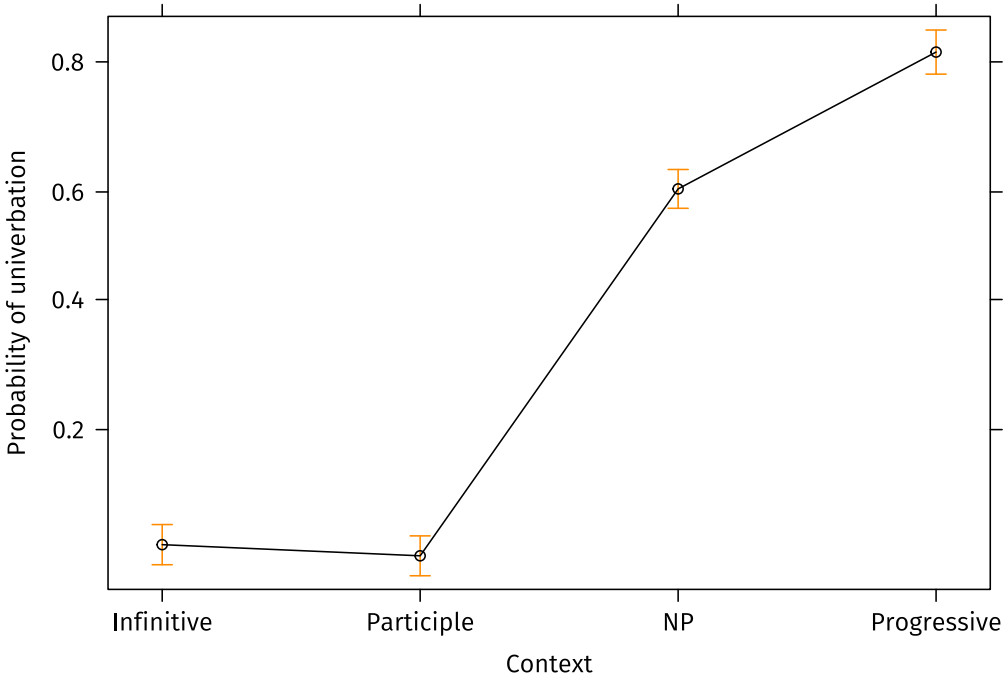

Figure 1: Effect plot for the regressor encoding the morphosyntactic context of the N+V unit in the GLMM modelling the corpus data.

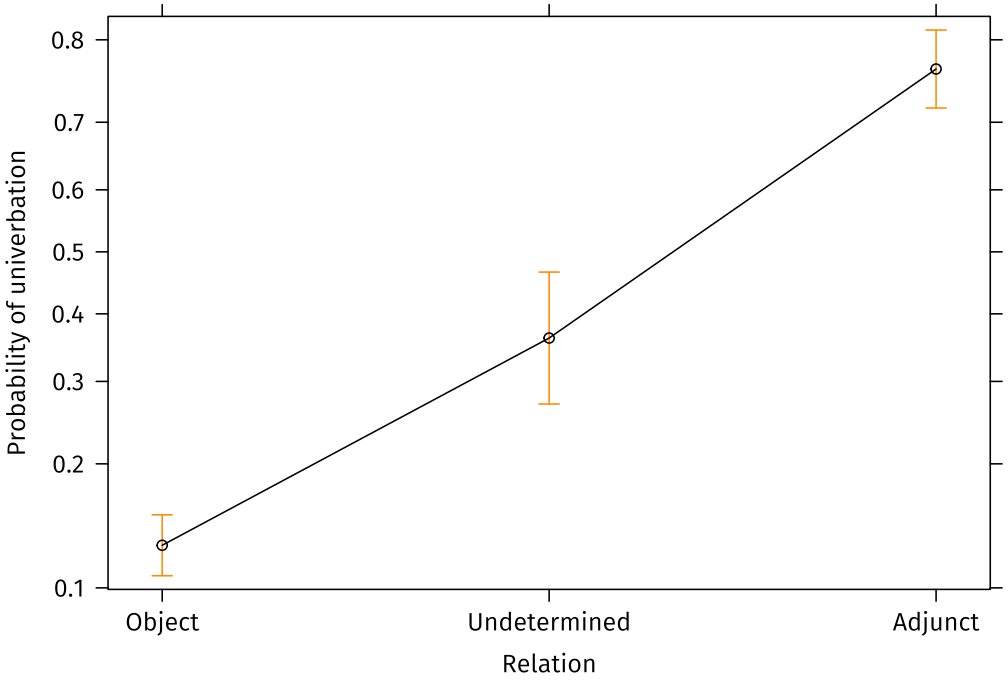

Figure 2: Effect plot for the regressor encoding the syntactic relation within the N+V unit in the GLMM modelling the corpus data.

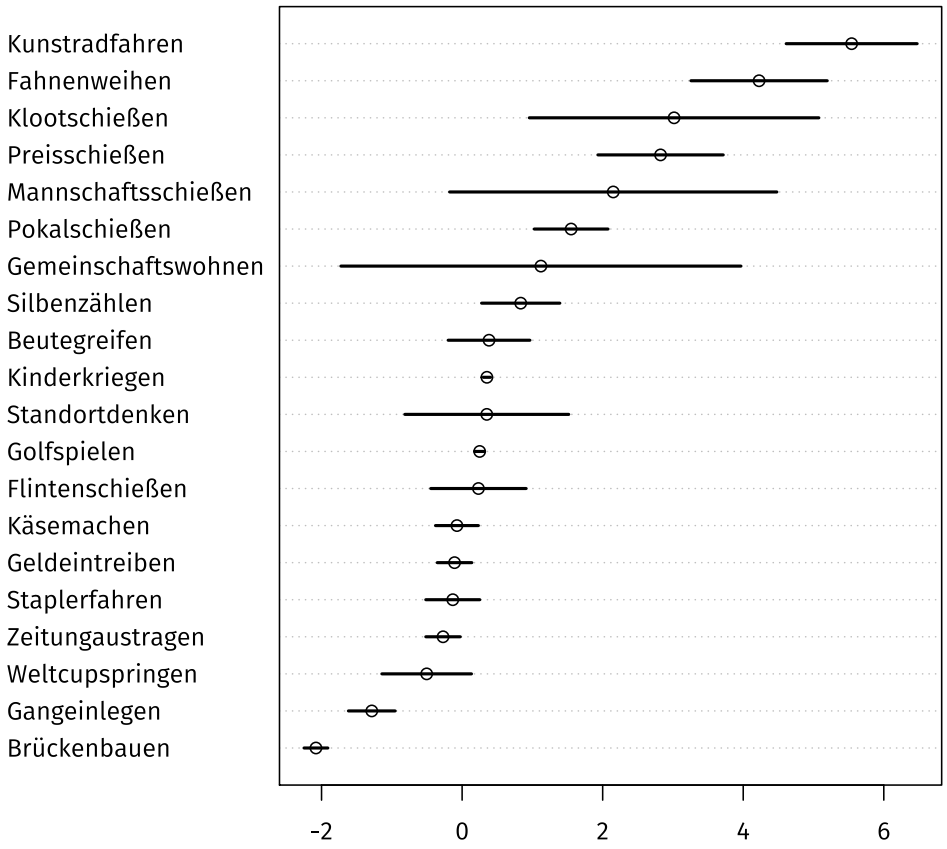

Figure 3: A random selection of conditional modes with 95% prediction intervals for the levels of the random effect in the GLMM modelling the corpus data.

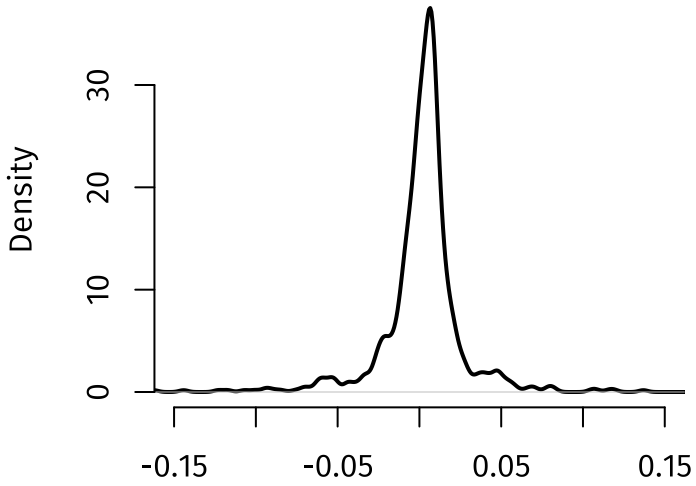

Figure 4: Density estimate of the distribution of the overall association scores (across all morphosyntactic conditions) with $n = 820$.

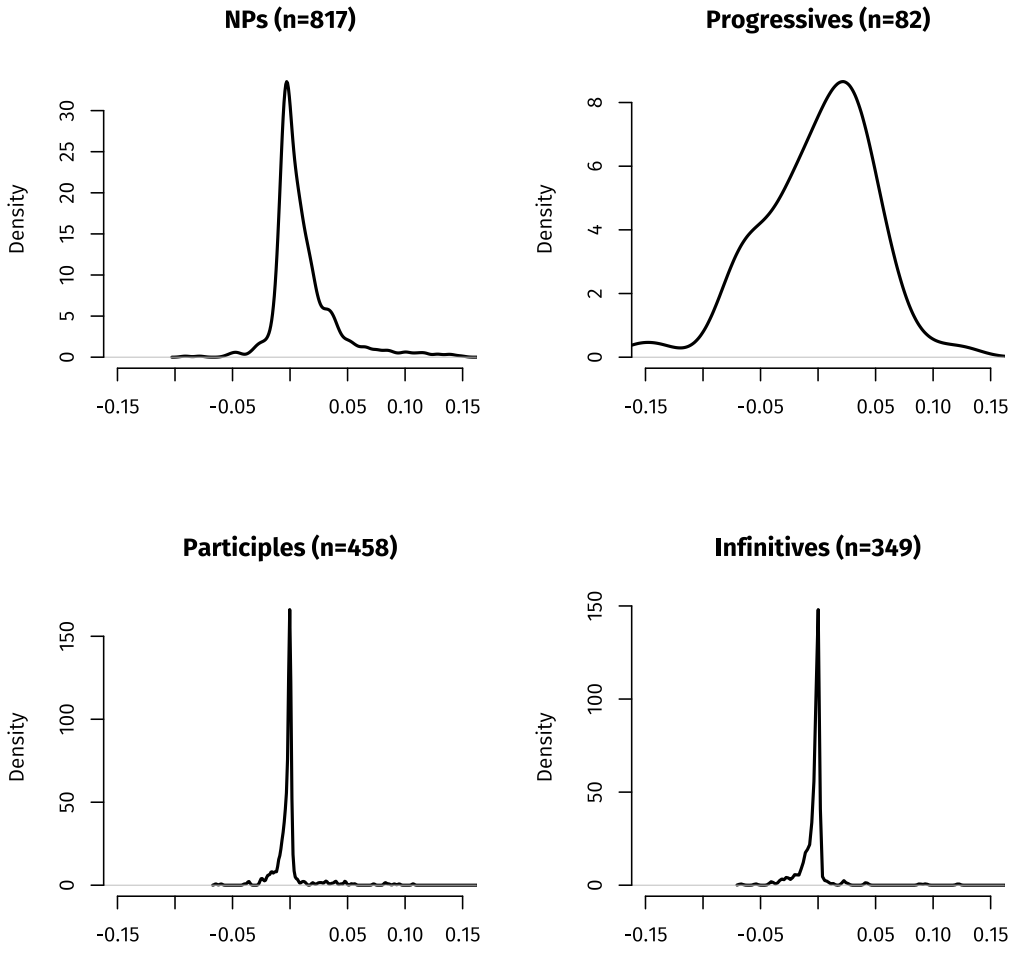

Figure 5: Density estimates of the distribution of the association scores in the specific morphosyntactic conditions. Because of some undefined scores the sample sizes n vary.

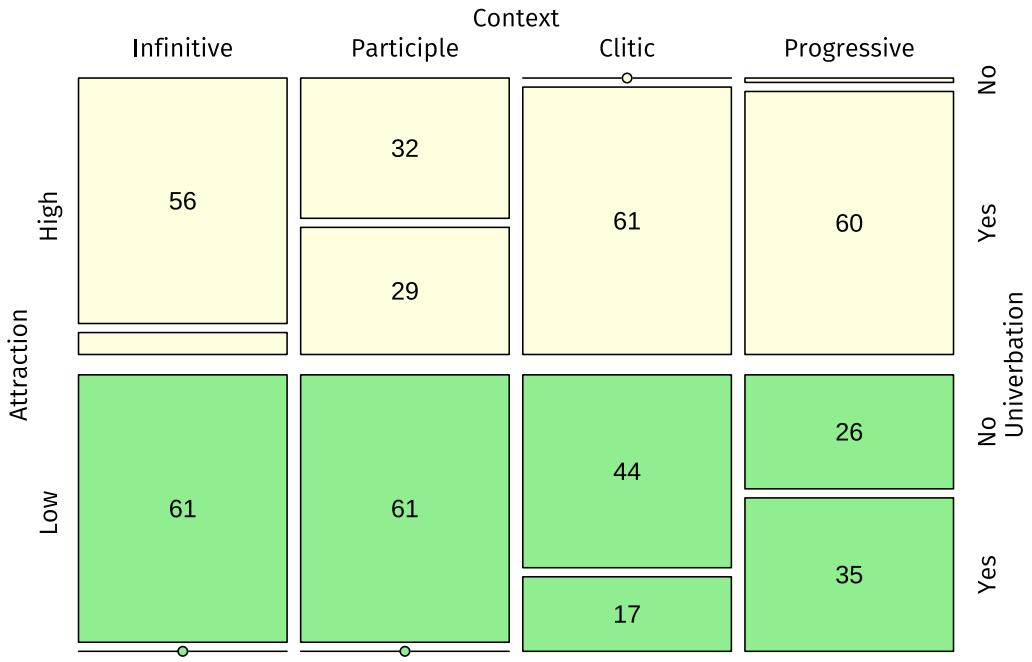

Figure 6: Mosaic plot of the responses in the production experiment (vertical right) grouped by the morphosyntactic context (horizontal) and the binned N+V unit's attraction strength calculated from the corpus (vertical left).

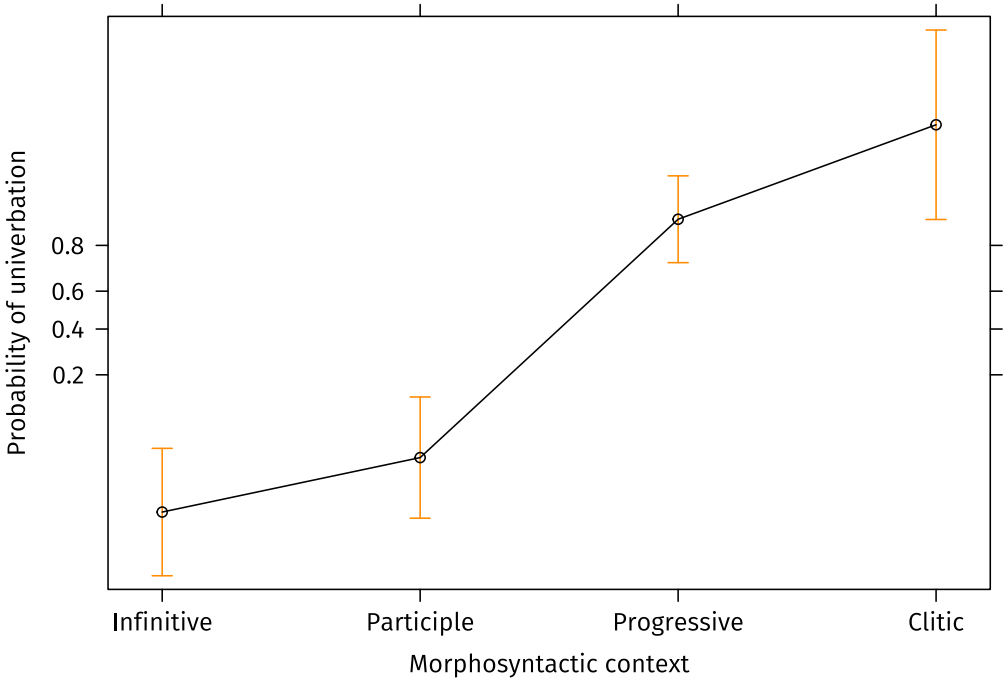

Figure 7: Effect plot for the regressor encoding the morphosyntactic context in the GLMM modelling the experimental data.

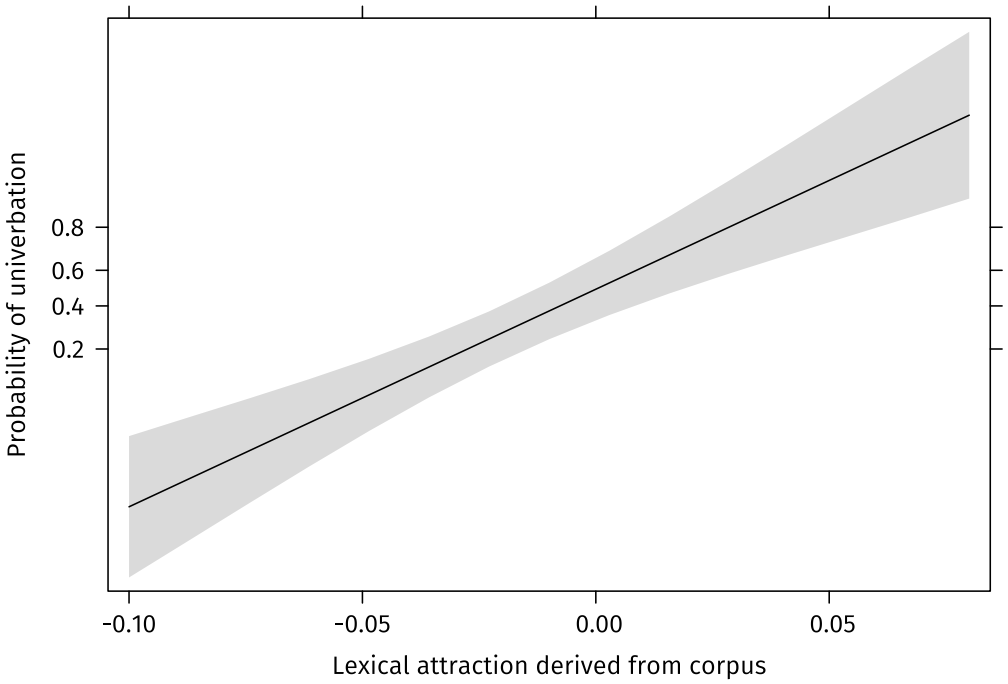

Figure 8: Effect plot for the regressor encoding the N+V unit's corpus-derived association with univerbation in the GLMM modelling the experimental data.

	Estimate	CI low	CI high
(Intercept)	-3.584	1.998	2.216
ContextParticiple	-0.084	-3.808	-3.360
ContextNP	2.682	-0.136	-0.032
ContextProgressive	3.714	2.643	2.722
RelationUndetermined	1.355	3.626	3.803
RelationAdjunct	3.114	0.881	1.829
LinkYes	0.351	2.792	3.438

Table 1: Coefficient table for the binomial GLMM modelling the corpus data with 95% confidence intervals. The response is for each N+V unit in each morphosyntactic context ($N_u = 3296$ derived from a total number of observations ($N_{obs} = 2,322,076$) the proportion of compound spellings among all of its spellings. Weighting was used to account for the bias in models on proportion data. Random effect for V+N lemma: $\text{Intercept} = 4.421$, $\text{sd} = 2.103$. The intercepts model the fixed effects Relation=Object and Link=No. Nakagawa & Schielzeth's $R^2_m = 0.519$ and $R^2_c = 0.999$.

V+N Unit	Association	Relation
Teilhaben	0.185	Object
Radfahren	0.180	Undetermined
Computerspielen	0.137	Adjunct
Zeitreisen	0.119	Adjunct
Skifahren	0.116	Adjunct
Autofahren	0.108	Undetermined
Probefahren	0.105	Adjunct
Bogenschießen	0.082	Undetermined
Schifffahren	0.080	Undetermined
Windsurfen	0.080	Adjunct

Table 2: Top ten V+N units with a strong tendency for univerbation.

V+N Unit	Association	Relation
Klavierspielen	0.009	Object
Theaterspielen	0.008	Undetermined
Filmemachen	0.008	Object
Autowaschen	0.007	Object
Zigarettenrauchen	0.006	Object
Haarewaschen	0.003	Object
Notenlesen	0.002	Object
Golfspielen	-0.001	Object
Wasserholen	-0.009	Object
Haareschneiden	-0.009	Object

Table 3: Top ten V+N units without any tendency for or against univerbation.

V+N Unit	Association	Relation
Gedankenmachen	-0.162	Object
Geldverdienen	-0.144	Object
Rechtgeben	-0.123	Object
Spaßhaben	-0.117	Object
Rechthaben	-0.107	Object
Kinderhaben	-0.101	Object
Zeitnehmen	-0.095	Object
Auftraggeben	-0.093	Object
Fehlermachen	-0.089	Object
Urlaubmachen	-0.085	Object

Table 4: Top ten V+N units with a strong tendency against univerbation.

3.4 Association strengths

4 Elicited production of noun-verb units in written language

5 Explaining the process of noun-verb univerbation

Acknowledgments

References

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

	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 $= 1.723$.

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.

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.

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.

units 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] In R notation, the specification is thus: `ProportionRelation+Link+(1|NVUnit)`.