Between syntax and morphology: German noun-verb units as reluctant compounds

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

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

Abstract We show that graphemic variation—at least in some writing systems—can be analysed in terms of grammatical variation given a usage-based probabilistic view of the grammar-graphemics interface. Concretely, we examine a type of noun + verb unit in German, which can be written as one word or two. We argue that the variation in writing is rooted in the units' ambiguous status in between morphology (one word) and syntax (two words). The major influencing factors are shown to be the semantic relation between the noun and the verb (argument or oblique relation) and the morphosyntactic context. In prototypically nominal contexts, a re-interpretation of the unit as a noun + noun compound is facilitated, which favours spelling as one word, while in prototypically verbal contexts, a syntagmatic realisation and consequently spelling as two words is preferred. We report the results of a large-scale corpus study and a controlled production experiment to corroborate our analysis.

Keywords: univerbation, usage-based theory, prototypes, corpus data, experiments, German

1 Introduction

1.1 Grammar, graphemics, and usage

In this paper, we apply usage-based grammar to a graphemic alternation phenomenon in German, arguing that properties of a probabilistic grammatical system can be inferred by examining written usage, i. e., from a

graphemic perspective. Usage-based Grammar (UBG; e. g., Bybee & Beckner 2009; Kapatsinski 2014; Tomasello 2003) is based on two core assumptions: (i) grammar is acquired using only general cognitive devices, (ii) grammar is determined only by general cognitive constraints and by the input. Since the input is always rife with variation, which is intrinsically probabilistic, a third assumption is crucial to some researchers: (iii) grammars are learned as probability distributions over possible forms, meanings, and form-meaning pairs. We embrace all three assumptions and apply them to a graphemic alternation phenomenon wherein certain noun-verb combinations in present-day German can be spelled as one word or two words. UBG is rarely extended to graphemics in such a way, but we view graphemics as a component of the language faculty on a par with components such as phonetics and phonology, and we consequently believe that graphemics should be viewed under the usage-based umbrella.[1] Much like the phono-component comprises regularities about how grammar is encoded in speech sounds, graphemics comprises similar regularities about how grammar is encoded in written symbols. Whether and how strongly the phono-component and the graphemics component are intertwined is determined by the type of script and the specific language. Ideograph-based writing systems like early cuneiform Sumerian (virtually complete separation) and phonographic writing systems like German (substantial overlap) represent extremes on a continuous scale (see Coulmas 1996 for an overview). For writing systems like German, the mappings to be learned include sounds to letters, parts of speech to spellings (e. g., capitalisation of nouns), syntactic categories to spaces and punctuation marks, etc. (Primus 2010).[2]

In UBG, corpus data (i. e., production data) are often used as evidence, sometimes cross-validated in behavioural experiments (see, for example, Arppe & Järvikivi 2007; Bresnan et al. 2007; Dąbrowska 2014; Divjak 2016; Divjak, Dąbrowska & Arppe 2016; Ford & Bresnan 2013; Pankratz & Van Tiel 2021; **!** **[**Schäfer 2018; Schäfer & Pankratz 2018**]**). This is justified because the probabilistic usage-based nature of the acquisition process as described at the outset should be reflected in the grammars of competent adult speakers/writers and not just in the acquisition process itself. Consequently, it should also be reflected in production data obtained from com-

[1] There is an intrinsic graphemic component in the huge body of work throughout linguistics based on popular corpora of written language. Although this is rarely acknowledged, We consider it important to focus on this component as well.

[2] Notice that a probabilistic view does not necessarily imply that there are no discrete or virtually discrete mappings like the one-to-one mapping of consonantal segments to letters in German. Cases of discreteness can always be seen as extremes in a probabilistic system.

petent adults, and we should be able to uncover the probabilistic mappings of lexical-grammatical categories to written forms from such data. We consequently use corpus data as well as data elicited in controlled experiments, both being forms of production data. However, there is a difference between using production data as evidence and assuming that they *directly* mirror cognitive reality. While it is generally assumed that corpora represent a valid source of data in cognitively oriented linguistics (e. g., Newman 2011), it is also known that there is no straightforward correspondence between corpus data and cognitive reality (e. g., Gries 2003; Dąbrowska 2016). What we hope to recover from corpus data are major abstractions learned by a majority of speakers, uncovering general cognitive principles that ideally go far beyond individual acquisition careers and idiosyncrasies of single languages.

A convenient framework to formulate such abstractions is Prototype Theory (Rosch 1973; 1978). As a cognitive theory of classification, it is compatible with probabilistic views since it allows for fuzzy category membership (e. g., Sutcliffe 1993; Murphy 2002: 11–16). Grammatical units can thus be modelled as belonging to multiple categories to different degrees or—in our case to be introduced immediately below—as alternating between a morphological and a syntactic realisation.[3] Prototype Theory is also intrinsically compatible with UBG as it assumes just a very general mechanism of classification whereby newly encountered objects are classified by similarity to a prototypical exemplar. In most versions of Prototype Theory, these prototypes are identified by (weighted) features or *cues*, and unseen exemplars are categorised depending on how many of those features they share with the prototype. We use Prototype Theory as a suitable framework in our analysis. Grammatical prototypes are mapped onto graphemic realisations (e. g., spellings), and the more strongly a unit matches the prototype, the more likely it is to be realised as the variant mapped to that prototype.

One caveat that is specific to graphemics needs to be mentioned before we proceed to the description of the concrete phenomena. The acquisition of the writing system involves explicit instruction and is thus more strongly imposed by prescriptive norms. However, we expect writers to learn grammar-graphemics mappings primarily from their realisations in the input, especially whenever the norm is unspecific or unclear, a situation which provides ideal test cases for our view of graphemics. Variation

[3] For applications of Prototype Theory in linguistics see, among many others, Divjak & Arppe (2013); Dobrić (2015); Gilquin (2006); Gries (2003); ! [Schäfer (2019)] . See Taylor (2003; 2008) for introductory overviews.

and alternations in the written input shape the acquired probability distribution, and conditioning factors are acquired to the degree that they can be retrieved from the type and the frequency of the input.[4] We are convinced that graphemics is a field in its own right which deserves attention in any grammatical/linguistic framework. See Berg (2016) for a compatible fundamental argument independent of a concrete grammatical framework.

1.2 Introducing N+V units in German

The alternation we are going to explore affects units containing a noun and a verbv, and these units alternate between a syntactic manifestation (where the noun combines with the verb via a syntactic mechanism) and a morphological one (where the noun is incorporated into the verb). We will argue that alternations in spelling provide evidence for the grammatical status of the instances of the construction. Simple examples not showing the alternation are given in (1).

(1) a. Remy fährt Rad.
Remy rides_{V} bike_{N}
Remy is riding/rides a bike.

b. Remy läuft Eis.
Remy runs_{V} ice_{N}
Remy is ice-skating/ice-skates.

(2) Remy läuft auf dem Eis.
Remy runs on the ice_{N}
Remy is running on the ice/is ice-skating.

In this construction, there is a noun N occurring in its bare form, which either corresponds to an argument of the verb V (normally in the accusative case) as in (1a) or to an adjunct of the verb V as in (1b), which would normally take the form of a prepositional phrase asn in (2).[5] As *eislaufen* is highly lexicalised, (2) is no longer a proper paraphrase of (1b), and it

[4] We have previously used a similar approach in, for example, ![Schäfer & Sayatz (2014; 2016)].

[5] Whereas singular indefinite mass nouns typically occur without an article in German (Vogel 2000: 471), this is the only frequent construction in German in which bare count nouns occur. However, there is a class of lexicalised light verb constructions where a bare noun occurs with a light verb, such as *Anklage erheben* 'indict', literally 'to raise indictment'. Like idiomatic expressions such as *Leine ziehen* 'get lost', literally 'to pull leash', they do not instantiate a productive pattern (Hentschel & Weydt 2003: 76, Stumpf 2015: 198). Consequently, we do not discuss them further.

has the more general meaning of just ‘walking on the ice’, which includes ‘ice-sakting’. However, many other V + N units are far less lexicalised, and they can all be transparently related to a paraphrase with a PP (see below).

We use the terms ‘argument relation’ and ‘oblique relation’ to refer to the semantic relations between the noun and the verb in the former and the latter case, respectively, following Gaeta & Zeldes (2017: 20). Oblique nouns occur without their usual preposition, and since the accusative case is only morphologically encoded on determiners (if at all) in German, the relation between the noun and the verb is never formally encoded in either case. Furthermore, the noun always acquires an unspecific generic reading: in examples such as (1a), *Rad fahren* (‘to ride bike’) refers to the concept of riding any bike, and the unspecific reading of *Rad* is obligatory, which is not the case for the English translations with the indefinite article.

German clausal syntax creates the conditions for the actual spelling alternation to occur; see (3).[6]

(3) a. Remy **fährt** gerade **Rad**.
Remy rides$_{\text{PRES}}$ right.now bike
Remy is riding a bike right now.

b. Yael weiß, dass Remy **Rad fährt**.
Yael knows that Remy bike rides$_{\text{PRES}}$
Yael knows that Remy is riding a bike.

c. Remy ist gestern **Rad gefahren**.
Remy is yesterday bike ridden$_{\text{PART}}$
Remy rode a bike yesterday.

d. Remy will **Rad fahren**.
Remy wants bike ride$_{\text{INF}}$
Remy wants to ride a bike.

e. Remy hat keine Lust, **Rad zu fahren**.
Remy has no motivation bike to ride$_{\text{INF}}$
Remy doesn’t feel like riding a bike.

f. Remy ist am **Rad fahren**.
Remy is at.the bike ride$_{\text{INF/NOUN}}$
Remy is riding a bike.

g. Remy singt beim **Rad fahren**.
Remy sings upon.the bike ride$_{\text{INF/NOUN}}$
Remy is singing while riding a bike.

[6] Further spelling variants for (3c) through (3h) will be discussed immediately below.

h. * Remy lobt das **Rad fahren**.
Remy praises the bike riding$_{\text{NOUN}}$
Remy praises the riding of bikes.

Such N + V units occur flexibly in all types of syntactic contexts: with finite verbs in verb-second order (3a), with finite verbs in verb-last order (3b), in the analytical perfect where the lexical verb takes the form of a participle (3c), in bare infinitives (3d), in infinitives with the particle *zu* (3e), in a progressive-like construction with the preposition *an* fusioned with the dative singular article *dem* to *am* where the infinitive is potentially nominalised (3f), and in regular prepositional phrases (3g). In (3h), the spelling of the N + V unit as two words is impossible, hence the asterisk. In this case, we can assume that the noun and a fully nominalised infinitive form a regular nominal compound.[7] The spelling as two words for (3f) and (3g) is not accepted by all native speakers.

In the examples (3c) through (3h), the noun and the verb occur in sequence without intervening material. In these cases, the noun and the verb alternate between the spelling as multiple words seen in (3) and spellings as one word shown in (4).

(4) c. Remy ist gestern **radgefahren**.
d. Remy will **radfahren**.
e. Remy hat keine Lust, **radzufahren**.
f. Remy ist am **Radfahren/radfahren**.
g. Remy singt beim **Radfahren/radfahren**.
h. Remy lobt das **Radfahren**.

In (4f) and (4g), additional variation is introduced in the form of upper-case and lower-case initials.[8] The compound with the nominalised infinitive in (4h) is fine if spelled as one word.

We call cases where a multi-stem unit is spelled as two words such as in (3) the 'disjunct spelling' and cases where a unit is spelled as one word as in (4) the 'compound spelling'. We see that N + V units potentially undergo graphemic *univerbation* in the form of compound spelling. Lehmann (2020: 206) calls univerbation "the union of two syntagmatically adjacent word forms in one". We follow this terminology and assume univerbation

[7] Infinitives in German can be routinely nominalised as an action noun (Gaeta 2010: 224, Dammel & Kempf 2018: 67, Werner, Mattes & Korecky-Kröll 2020: 172–174).
[8] In German, all nouns are capitalised anywhere in a sentence (Pauly & Nottbusch 2020: 1).

to be the directly observable phenomenon, i. e., compound spelling of adjacent words that could potentially also be used in disjunct spelling or were historically used in disjunct spelling. Historically, univerbation is a gradual process, and it can thus be a strongly probabilistic phenomenon due to the slowly changing grammatical and lexical system. However, univerbation per se is not necessarily the result of a regular grammatical pattern or process.[9] Thus, a major aim of this paper is to show whether and how the univerbation of N + V units in German is based on an established morphological prototype construction wherein a noun is incorporated into a verb, forming a new verb expressing a new event concept.

We will argue that such a morphological construction exists, but that the alternative syntactic construction remains available to speakers because N + V units have properties of both morphological as well as syntactic prototypes. In Section 2, we lay the theoretical and descriptive foundations. We then present a large-scale corpus study and an elicitation experiment in Sections 3 and 4, exploring our particular hypotheses about N + V units. We conclude with a summary, further interpretation and discussion in Section 5.

2 Theoretical and descriptive background

In this section, we lay the theoretical and descriptive foundations of our analysis of the alternation described in Section 1. First, we clarify the status of spaces as syntactic boundaries in German in Section 2.1. Then, we discuss previous analyses of N + V units and their spelling, followed by the formulation of our predictions for the empirical studies, in Section 2.2.

2.1 Spaces, words, and univerbation

As explained in Section 1, we use graphemic evidence from corpora and controlled experiments, and we argue that it indirectly allows us to draw conclusions about writers' cognitive grammars. More specifically, we assume that compound spellings of N + V units indicate that writers conceive

[9] For Gallmann (1999: 294), univerbation is a diachronic process wherein a complex syntactic unit is reanalysed as a simplex syntactic unit. Jacobs (2005: 107) regards graphemic univerbation as not rooted in a morphological process as they are not paradigmatic (no *Reihenbildung*). Lehmann (2020: 209) is closest to our position, as he regards "univerbation as a gradient process which displays phases of weaker and stronger univerbation". According to him, it marked by the loss of morphological boundaries and by phonological fusion.

of those units as single syntactic words, whereas disjunct spelling indicates that they conceive of the unit as two syntactic words. Therefore, we briefly introduce the status of the space in German writing and how it pertains to N + V units.

German writing uses an alphabetic script with a strong correlation between underlying phonological forms (the phonemic level) and characters (graphemes). A common fundamental principle of such scripts is the separation of syntactic words by spaces (Jacobs 2005: 22). Also, stems and their affixes are never separated from one another, which reinforces the status of the space as a demarcation of syntactic words.[10] These factors facilitate the reader's ability to decode the sequence of syntactic words, and they constitute a crucial principle in the encoding and conventionalisation of meanings associated with word forms (Jacobs 2005: 22).

Unlike in English, German has regular compound spelling of syntactic words comprising more than one stem, especially for the case of the highly productive noun + noun (N + N) compound pattern (Fuhrhop 2007: 182, Jacobs 2005: 34, Section 2.2 below), for which compound spelling is the dominant graphemic realisation. However, there is a heterogeneous group of multi-word constructions for which only a tendency towards compound spelling can be observed (Szczepaniak 2009: 95, Wurzel 1998: 335). As opposed to N + N compounds, these constructions typically consist of words with different parts of speech, such as *mithilfe* (*von*) ('with the help (of)') from *mit der Hilfe* (*von*) or *zuhause* ('at home') from *zu Hause*.[11] For such cases, Lehmann (2020: 206) posits a "downgrading of a syntactic to a morphological boundary" between the two words. When writers use compound spelling in these cases, they choose to encode the construction as a single word with a morphological boundary instead of a sequence of words with a syntactic boundary. If many speakers consistently make this choice over a significant period of time, the unit might become lexicalised as a single word (Lehmann 2020: 212). Until such a diachronic process is complete and one of the spellings has become clearly dominant, the item alternates between a syntactic and a morphological realisation. For many of these con-

[10] There is a class of verbal particles which does not follow this principle. Verbs like *aufessen* ('eat up') formed from a verb stem (*essen*) and a prefixed particle (*auf*) are spelled as one word when they are adjacent in verb-last order, but they are separated in verb-second order where the verb is moved to sentence-second position and the particle remains in sentence-last position through obligatory long-distance movement (see Hoberg 1981 for an account of German clausal and sentential syntax).

[11] Normative approaches as well as individuals display a lot of variation with respect to at least some of those constructions (cf. below).

structions, this is the case both in non-standard as well as standard written German, albeit to different degrees.[12]

N + V units with different affinities towards compound spelling like *Rad fahren* ('bike riding', often also spelled *radfahren*) and *eislaufen* ('ice skating', infrequently also spelled *Eis laufen*) represent different levels of diachronic re-conventionalisation as single words.[13] This indeterminacy means that speakers have both the syntagmatic realisation (disjunct spelling) and the morphological realisation (compound spelling) in their graphemic input, which subsequently leaves them with quite a free choice to be made based on how a concrete item is classified according to their individual grammar. It is the task of usage-based probabilistic graphemics to uncover factors influencing such decisions and decode the principles at work in speakers' internal grammar by analysing their writing habits (see ![Schäfer & Sayatz 2016]).

2.2 The status of noun+verb units in German

In Section 1, we showed that N + V units alternate between compound spelling and disjunct spelling when they occur in sequence. In this section, we explain why the existence of this alternation is not surprising considering the morphosyntactic system of German. Furthermore, we argue that in each concrete case where an N + V unit is written, the strength of the tendency towards either compound or disjunct spelling can be derived from the overall syntactic and morphological patterns available in present-day German. These patterns are shown to have prototypical properties which are matched more or less well by individual N + V units and their syntactic contexts, which leads to either compound or disjunct spelling being the preferred re-

[12] We are not aware of any published research systematically comparing the alternation tendencies in standard and non-standard written German.

[13] The orthographic norm is notoriously unstable with respect to N + V units, which contributes to their unclear status. Before the significant reform of the orthographic norm in 1996, both *radfahren* and *eislaufen* were supposed to be spelled as one word. After the reform, both units were supposed to be written as two words (*Eis laufen* and *Rad fahren*). After a revision of the reform in 2006, *eislaufen* was again supposed to be spelled as one word, whereas *Fahrrad fahren* was supposed to be spelled as two words exclusively (Primus 2010: 32, Eisenberg 2020: 356). From experience, we know that the norm is often not adhered to, and the data presented in Sections 3 and 4 strongly corroborate this experience.

alisation.[14] The hypotheses put forward here are then evaluated empirically in Sections 3 and 4.

To this end, we will shed some light on the productive N + N compound construction in Section 2.2.1 before turning to N + V units as reluctant compounds in Section 2.2.2. We sum up our arguments and derive our hypotheses for the empirical studies in Section 2.2.3.

2.2.1 N+N compounds

For an N + V unit to systematically undergo graphemic univerbation (i. e., a downgrading from a syntactic to a morphological construction in the sense of Lehmann 2020: 206), it must resemble one or more established morphological constructions closely enough to be classified as an instance of such constructions itself.[15] We posit that this follows from the assumed underlying learning mechanisms under a usage-based perspective. The prototypical and arguably the only fully productive morphological construction combining more than one stem in German is noun + noun (N + N) compounding, to which we turn now in some detail.[16] German N + N compounds instantiate a proper morphological construction. Syntactically, nothing can intervene in between the two stems of the compound, and they cannot be rearranged. With minor exceptions (often exaggerated in normative discussions), they are also inseparable graphemically, i. e., they are always written as one word (Scherer 2012: 57–60). Furthermore, they are always head-final, mostly determinative, and they allow recursive formation wherein an N + N compound enters into another N + N compound, resulting in [[N + N] + N] or [N + [N + N]] structures (Fleischer & Barz 2012: 13). Some examples are given in (5) and (6), the latter being recursively formed from the former.[17]

[14] Hüning (2010) describes a similar alternation of Adjective + Noun constructions in Dutch and German. He, too, argues that the respective constructions alternate between a syntactic and a morphological realisation, and he uses analogy to existing categories to explain the alternation. While we opt for a prototype description, Hüning's view is still based on the same underlying assumptions as ours.

[15] Random isolated univerbations like *zuhause* 'at home' from *zu Hause* are not systematic in this sense. They are merely the result of idiosyncratic diachronic developments.

[16] Adjectives also enter compounds as the head, such as in *feuerrot* 'red like fire', literally 'fire red'. However, this pattern is much less productive than N + N compounding, and we do not discuss it here. See Simunic (2018: 136) on the productivity of N + A compounds.

[17] If necessary, we present compound spelling with a minimal analysis of the morphological structure. Affixes are separated from stems by hyphens, and lexical stems are separated from each other by a period. Within compounds containing more than two stems, structure is shown using square brackets.

(5) Haus.tür
house.door
front door

(6) Haus.tür.schlüssel
[[house.door].key]
key to the front door

The semantic relation between the first noun (N_1) and the second noun (N_2) is highly unspecific, rendering many compounds semantically ambiguous unless they are strongly lexicalised (Klos 2011: 252).[18] The historic development of the stable N + N compound construction was furthered during the Early New High German period (approximately from the 14th to the 17th century AD) by a syntactic change.[19] The dominant pattern of nominal attribution within the NP had been a prenominal genitive as in the now obsolete (7), which swiftly changed to a postnominal genitive as in (8).

(7) † des Hauses Tür
the_{Gen} $\text{house}_{\text{Gen}}$ door
the door of the house

(8) die Tür des Hauses
the door the_{Gen} $\text{house}_{\text{Gen}}$
the door of the house

To the extent that prenominal attribution in syntax became more and more obsolete, the prenominal position was used to establish the highly productive morphological construction of N + N compounds as in (5) (see Nübling et al. 2017: 132, Schlücker 2012), which showed a tendency to be written in compound spelling very early on (Dücker & Szczepaniak 2017: 34–36). The N + N compound construction is semantically at least as unspecific as the syntactic genitive construction to which it is diachronically related (Eisenberg 2020: 239). Its recursive application is virtually unrestricted (Wurzel 1994: 504). N_1 and N_2 are just concatenated as bare stems in most cases, but there are also so-called linking elements, which

[18] Obviously, once they are strongly lexicalised, they cannot help to establish a more canonical type of semantic relation between N_1 and N_2, either, simply because lexicalised compounds are often intransparent to the language user (Klos 2011: 59), such as *Kammerjäger* ('pest controller', literally 'chamber hunter').

[19] Nübling et al. (2017: 130) find that the prenominal genitive begins to give way to the postnominal genitive in the 13th century. Around 1500, already 53% of the genitives are postnominal, rising to 64% at around 1700.

are sometimes positioned in between the stems.[20] Diachronically, linking elements arise from diverse sources, but the overall pattern of inserting them is related to the former morphological marking in prenominal genitives (Nübling et al. 2017: 55–57).

N + N compounds as described in this section are clearly the prototype for morphological constructions combining more than one stem in German. In the next section, we show how N + V units deviate from this prototype, and how this keeps them from univerbating fully and stably.

2.2.2 N+V units as reluctant compounds

In this section, we argue that N + V units are *reluctant compounds*. Some factors pull away from compounding and univerbation, other factors pull them towards it. While in principle the morphological N + V construction (as a kind of compound written as one word) has existed for centuries, we show that it remains in competition with a syntactic construction because it does not fit the morphological compounding prototype very well. At the same time, we suggest why morphological compounding and consequently the spelling as one word (univerbation) become the preferred realisation under the right circumstances.

Most likely, full compounding of N + V units requires conversion of the verbal head to a noun, making the result an N + N compound via conversion. As opposed to compounding with proper nominal heads (as discussed in the previous section), compounding with verbal heads is not a productive pattern in German.[21] A major difference to N + N compounds is the fact that N + V units are usually not inseparable, as was already shown in Section 1. There can be intervening material in between the noun and the verb in some contexts, namely the infinitival particle *zu*. This particle is, however, virtually part of the verbal word form and does by no means prevent univerbation (see example 4e, where *radzufahren* is spelled as one word instead of *rad zu fahren*). Furthermore, in a verb-second sentence, the noun may remain at the end of the sentence in a structure reminiscent of particle verbs (Fortmann 2015: 603), see (3a). This fact alone means that N + V units do not fit the compounding prototype anywhere near perfectly. This

[20] A recent large-scale study (Schäfer & Pankratz 2018: 339) showed that 60% of all N + N compound types have no linking element, whereas 40% have one of several possible linking elements.

[21] Günther (1997) counts roughly 400 lexicalised N + V compounds in Muthmann (1988) (see also Eisenberg 2020: 245).

likely introduces great resistance in speakers to classify them as compounds and use compound spelling.

Another major difference between N + N compounds and N + V units is that the morphological N + V construction is not recursive. Nominalised N + V units marginally occur as N_1 in $N_1 + N_2$ compounds (contrary to claims by Fuhrhop 2007: 54), as in (9).[22] However, an N + V unit cannot function as the verbal head in another N + V unit (i. e., a [N + [N + V]] structure) as illustrated in (10). Native speakers will readily acknowledge that such constructions are absurd.

(9) a. Energie.spar.messe
[[energy.save].fair]
trade fair for products useful in saving energy

b. Endlager.such.gesetz
[[final storage.search].law]
law about the search for a permanent repository for nuclear waste

c. Feuer.lösch.boot
[[fire.extinguish].boat]
fire-fighting boat

(10) * Rad.fahr.mach-en
[[bike.ride].make-INF]
make bike riding

We posit that the lack of core properties of prototypical (productive) German compounding constructions (separability, potential reordering, lack of recursive application) is a major factor in keeping the formation of N + V units from establishing a fully productive morphological compounding construction, thus keeping it from reliably requiring graphemic univerbation.

Another noticeable difference between the N + N and the N + V construction is the specificity of the internal relation. While the relation in N + N compounds is quite varied, unspecified, and unspecific (see Section 2.2.1), there are only two possible relations within N + V units as we saw above, and these relations are determined by the verb's argument structure. The relation is always either an argument relation, as in (11), or an oblique

[22] The examples in (9) are attested and taken from the DECOW16B web corpus (see Section 3.1). Their document frequencies are 218 for *Energiesparmesse*, 416 for *Endlagersuchgesetz*, and 414 for *Feuerlöschboot* in a corpus of 17.1 million documents. The document frequency is the number of documents the lemma occurs in, not counting multiple occurrences within each document.

relation, as in (12). These examples show that there is a syntactic paraphrase for N + V units (where, for the oblique relation, the noun occurs in a prepositional phrase that is an adjunct to the verb).[23]

(11) a. Kim will (*eine Tasse) teetrinken.
Kim wants (a cup) tea drink
Kim wants to drink tea.

b. Kim will (eine Tasse) Tee trinken.
Kim wants (a cup) tea drink
Kim wants to drink (a cup of) tea.

(12) a. Kim will die Corvette probefahren.
Kim wants the Corvette test.drive
Kim wants to test-drive the Corvette.

b. Kim will die Corvette zur Probe fahren.
Kim wants the Corvette to the test drive.
Kim wants to test-drive the Corvette.

This relation can be paraphrased for almost all N + V units. The exceptions are the rare cases which have been lexicalised so fully that their meaning has changed significantly. This means that the morphological construction marked by graphemic univerbation remains in competition with a syntactic construction with distinct syntactic words separated by spaces in writing. This competition between a morphological construction and a syntactic construction was pointed out with varying terminology by—among others—Fleischer & Barz (2012: 12), Schlücker (2012: 13), and Morcinek (2012: 88). Whereas the alternative syntactic construction for N + N compounds (the prenominal genitive) disappeared within a relatively short period of time, the ambiguity between syntax and morphology of N + V units remains intact. This is true even though univerbation of N + V units with an argument relation dates back to Middle High German (*lobpreisen* 'praise', literally 'to praise compliment') and even Old High German (*hals-werfōn* 'turning around', literally 'to throw neck'); see Wurzel (1994: 517), Wurzel (1998: 334). For N + V units with an oblique relation, Morcinek (2012: 89) notices that dictionaries from between 1750 and 1993 list novel N + V units with an oblique relation with increasing frequency. For centuries or perhaps even more than a millennium, N + V units have been co-existing in syntax and morphology. We assume that the stable availability of an alternative

[23] Pragmatically, these paraphrases might often be subject to blocking because of the availability of the N + V construction. However, this does not make them syntactically or semantically unacceptable.

syntactic realisation is yet another major factor in preventing N + V formation from becoming a more clearly morphological construction in language users' cognitive grammars, making N + V units *reluctant compounds*.

So far, we have focussed on the factors that speak against the compounding of N + V units. We still have to show why and under which conditions true compounding and subsequent univerbation (including potential V-to-N conversion of the head and subsequent graphemic univerbation) are preferred. As already mentioned, the morphological realisation of N + V units is a type of noun incorporation. N + V units are usually seen as the only cases of potential incorporation in Modern German (Eisenberg 2020: 245), which is why we argued above that argument and oblique relations are prototypically realised syntactically in the form of objects and adjuncts. According to Mithun (1984: 848), incorporation is "a particular type of a compounding in which a V an N combine to form a new V".[24] As Mithun (1984: 848–849) points out, incorporation happens when the verb denotes a new and independent event concept in combination with the incorporated noun, and the semantics of the event is determined by the previous syntactic relation between the noun and the verb. Typically, the noun loses its referential autonomy as well as its specificity, and it acquires a generic reading, which is indeed the case for N + V units. In sentences like (13), no specific bike is referenced, and *radfahren* refers to the whole concept of riding any bike. This is true for both compound and disjunct spelling.

(13) Friedel kann radfahren/Rad fahren.
Friedel can bike.ride
Friedel knows how to ride a bike.

As a result of the semantic degradation of the noun, it looses its modifiability (also regardless of spelling), as illustrated in (14).

(14) * Friedel kann schnelles Rad fahren.
Friedel can quick bike ride
Friedel knows how to ride a quick bike.

[24] From Mithun's types of noun incorporation, German N + V units clearly represent type 1 *lexical compounding*. Since this is highly straightforward and homogeneous, we do not discuss the literature on the typological classification of incorporation further. Also, we do not take up the typological discussion on whether noun incorporation is a syntactic or a morphological process (e. g., Haugen 2015: 414–421, Mithun 2000: 923–925). We argue that N + V units in German clearly undergo a word-formation process when the noun is fully incorporated into the verb, and it's the alternation between syntax and morphology that is of particular interest to us. Semantically, both cases (syntactic and morphological combination) have the same properties which are typical of incorporation.

Such losses of referential autonomy and syntactic combinatorics are referred to as 'noun stripping' by Gallmann (1999: 287). The loss of specificity and referential autonomy as well as the acquisition of a generic reading are part of the semantics of the N + V construction (see also Gallmann 1999: 287, Bredel & Günther 2000: 108, Eisenberg 2020: 354). Functionally, the construction exists in order to express the new event concept which requires the generic/unspecific reading of the noun. Thus, the noun has the properties typical of nouns that are subject to incorporation of the lexical compounding type. In contrast with the factors that prevent univerbation and make N + V units reluctant compounds, this is a factor that we assume to facilitate their univerbation.

In the next section, we will summarise the factors that influence the tendency of N + V units to undergo univerbation. These factors will then be analysed in the empirical work to be reported in Sections 3 and 4.

2.2.3 Conclusions for the empirical studies

In the previous sections, we have laid out a theory of the factors favouring the univerbation of N + V units. In this section, we summarise and describe the concrete effects that we expect to see in written production data based on our overall usage-based framework (see Section 1) and our theoretical assessment of N + V units. These effects are then examined empirically in Sections 3 and 4.

In general, greater similarity to the N + N compound prototype and reduced competition from the full syntactic realisation are expected to favour univerbation. An important cue for this prototype is a strongly nominal morphosyntactic context. Concretely, when the unit is embedded in an unambiguously verbal syntagma (e. g., when the V head is an infinitive dependent on a modal verb or a participle dependent on an auxiliary), we expect a low tendency towards univerbation.[25] However, when the unit occurs in a strongly nominal syntagma (e. g., when the head is a fully nominalised head of an NP with a determiner), we expect a high tendency towards univerbation due to an accessible interpretation as an N + N compound.

The second important cue is the internal semantic relation. N + V units with an oblique relation stand in weaker competition with a syntactic realisation compared with those that have an argument relation. N + V units

[25] Notice that a finite V head would also create a prototypically verbal context. However, in these cases, the N and the V are always realised discontinuously, and univerbation is not an option.

with an oblique relation would need more explicit marking with a preposition in the alternative (unambiguously syntactic) realisation.

The univerbation of individual N + V units also involves very long-term diachronic processes of lexicalisation (see Section 2.2.2). While the individual lexicalisations are likely driven by the prototypicality effects described here, individual units might have progressed farther than others on the lexicalisation path. Furthermore, when the compositional meaning of individual N + V units becomes less accessible, univerbation might be favoured as the semantics of the unit becomes more holistic. Since philological investigations into the fate and semantics of each individual N + V unit are not feasible due to their sheer number, we will capture such individual tendencies numerically by comparing the frequencies of the units with or without univerbation in current usage (collexeme analysis). The results of this analysis will be used to control for such effects in the experiment reported in Section 4.

Finally, different speakers should be expected to have individual tendencies due to the variance in their input and in their compliance with normative advice. While individual variation can rarely be controlled in corpus studies due to the lack of metadata identifying individual writers, it should be controlled and/or analysed in behavioural experiments.

At any rate, under a probabilistic usage-based view of language, all these factors are expected to influence univerbation non-deterministically. Even in cases where all factors favour a realisation with univerbation, writers might sometimes spell it without univerbation and vice versa. However, we expect such cases to be rarely found in usage data if the hypotheses put forward here correctly describe reality. Our statistical models will be chosen appropriately for this assumption.

3 Analysing the usage of noun+verb units

In this section, we apply two quantitative methods to analyse the univerbation of N + V units using corpus data. We motivate our choice of corpus and describe the sampling and annotation procedure in Section 3.1. We perform exploratory analysis using association measures in Section 3.2 in order to gauge the individual tendencies of N + V units to undergo univerbation in written language usage. Finally, the results of estimating the parameters of a multilevel model explaining the variation in the univerbation of N + V units are reported in Section 3.3.

3.1 Choice of corpus, sampling, and annotation

As a first step, we adopted a data-driven approach in order to find close to all N + V units in contemporary written usage. In a second step, we counted their occurrences in compound and disjunct spelling in the relevant morphosyntactic contexts enumerated in Section 2.2.3: fully nominalised as the heads of noun phrases, in *am* progressives, as participles in analytical verb forms, and as infinitives in a range of verbal constructions.

Clearly, a large corpus with rich morphological and morphosyntactic annotations containing texts written in a broad variety of registers and styles (including ones written under low normative pressure) was required. We chose the DECOW16B corpus (Schäfer & Bildhauer 2012) because it fulfils all the aforementioned criteria.[26] 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 several 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 corpus searches of roots within nominal compounds. For example, we could query compounds with a deverbal head such as *Zeitnehmen* ('time taking').

The list of actually occurring N + V units was obtained by querying for compounds with a nominal non-head and a deverbal head.[27] The rationale behind this approach is that any N + V unit of interest should occur at least once in compound spelling as a fully nominalised compound. Since this step relied on automatic annotation already available in the corpus, the results contained erroneous hits which we removed manually. The resulting list contained 819 N + V units.[28]

In the second step, we created lists of all relevant inflectional forms of the verb in each N + V unit and used these to query all possible compound and separate spellings (including variance in capitalisation) of each of the 819 N + V unit types. In total, 28,665 queries were executed to create the

[26] https://www.webcorpora.org

[27] See the scripts available under the following DOI for concrete queries and further details: TO BE INCLUDED IN THE ACCEPTED VERSION.

[28] Notice that three highly frequent N + V units were excluded because they could be considered outliers, as they have fully undergone lexicalisation and are vitually always used in compound spelling. They are *Teilnehmen* 'to take part', *Maßnehmen* 'take measure', and *Teilhaben* 'have part' (meaning 'to participate'.

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 tokens.

For each N + V unit in the sample, the following variables were annotated automatically: (i) the verb lemma, (ii) the noun lemma, and (iii) 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. The codes used in clear-cut cases were *Argument* (441 units) and *Oblique* (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 correspond to the paraphrase either in (15a) or in (15b).

(15) a. [den Daumen]$_{\text{NP}_{\text{Acc}}}$ lutschen
the thumb suck

b. [am Daumen]$_{\text{PP}}$ lutschen
on the thumb suck

The data thus obtained were analysed in two ways. First, we report the results of a collexeme analysis in Section 3.2, which quantifies how strongly individual N + V units tend to be written as one word or two words. The association strengths were used as a covariate in the analysis of the experiment reported in Section 4. Second, in Section 3.3 we report a full statistical model of the alternation.

3.2 Results 1: Association strengths

In this section, we report an analysis of the item-specific affinities of N + V units towards univerbation. The method we use is similar to collocation analysis (Evert 2008 for an overview) and stems from Collostructional Analysis (Stefanowitsch & Gries 2003). More specifically, the method is called *collexeme analysis* (Stefanowitsch & Gries 2009).[29]

Our goal was to quantify how strongly each N + V unit tends towards univerbation vis-a-vis all other N + V units. Thus, we need to compare the counts of cases with and without univerbation of the unit in question with the total counts for all other N + V units. Such comparisons must be made relative to the overall number of the specific N + V unit as well as the num-

[29] See also ![Schäfer & Pankratz (2018)] and ![Schäfer (2019)] for similar uses.

ber of all other units. The counts needed for each N+V unit are nicely summarised in a 2×2 contingency table as shown in Table 1.

	Compound spelling	Disjunct spelling
Specific N+V unit	c_{11}	c_{21}
All other N+V units	c_{21}	c_{22}

Table 1: 2×2 contingency table as used in the calculation of the strengths of the associations of N+V units with univerbation.

With these counts, we are able to quantify how strongly the proportions in the first row differ from those in the second row, and there is a range of statistical measures for that. For example, one could use odds ratios or effects strengths from frequentist statistical tests.[30] We chose Cramér's v derived from standard χ^2 scores ($v = \sqrt{\chi^2/n}$). The v measure quantifies for each individual N+V unit how strongly the counts (cells c_{11} and c_{21}) deviate from its counts that we would expect if there were no difference between this unit and all other N+V units (cells c_{21} and c_{22}) with respect to their tendency to univerbate. Since Cramér's v is always in the range between 0 and 1, it allows us to compare analyses where the samples differ. 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.

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 1.[31] The graph shows the distribution of the attraction strengths for N+V units with argument and oblique relations separately. While there is variation in both directions in both cases, the argument relation tends more towards disjunct spelling (lower/more negative scores), and the oblique relation favours compound spelling more (higher/more positive scores). The number of units close to 0 (i. e., without a clear tendency) is notable with the argument relation. Examples include *Zeitreisen* ('time travel') with an oblique relation

[30] P-values from frequentist statistical tests are measures of evidence, 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 often used with measures of effect strength (Gries 2015).

[31] As expected, it approximates a scaled symmetric χ^2 distribution with $df = 1$ squashed between -1 and 1.

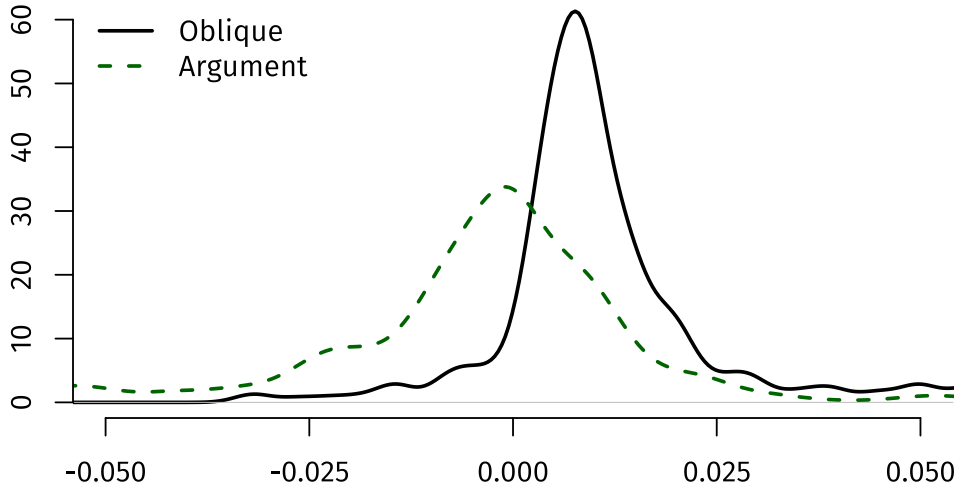

Figure 1: Density estimate of the distributions of the 819 association scores by the two semantic relations; the x-axis was truncated at -0.05 and 0.05 where the curves are essentially flat.

and a strong tendency towards univerbation (0.125), *Fehlermachen* ('mistake make') with an argument relation and a strong tendency against univerbation (-0.088), and *Haareschneiden* ('hair cut') with an argument relation and no clear tendency towards or against univerbation (-0.007). These results will be corroborated by the analysis in Section 3.3, and we will use the scores to control for item-specific tendencies in the experiment in Section 4.

3.3 Results 2: Multilevel model

In this section, we present the parameter estimates (and predictions of conditional modes) for a binomial multilevel model (or generalised linear mixed model, GLMM) which models the relevant factors influencing writers' choice of the compound and the separate spelling.[32] The results of the method used in Section 3.2 and the GLMM presented here converge. However, the GLMM has a more standard interpretation and allows for finer-grained data analysis. Also, it has long been accepted that combining several methods strengthens the analysis when the results converge (e. g., Arppe & Järvikivi 2007).

[32] See ! [Schäfer (2020)] for an overview of the method and our philosophy in modelling.

Given the grand total of 2,246,886 observations in the sample (see Section 3.1), we will completely refrain from an interpretation of the GLMM in terms of frequentist inferential statistics. For samples of such magnitude in data-driven approaches, frequentist significance tests are the wrong tool. 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 and predictions. 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 2.2, 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, 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 the spellings of the N + V unit. In the input data provided to the estimator, the response variable was thus a vector of 819 proportions, one for each N + V unit.[33] We specified four regressors. The only first-level (or observation-level) fixed effect regressor is the morphosyntactic context (a four-way categorical variable). We decided to break down the possible morphosyntactic contexts into four types:

i. adjunct *zu* infinitives (see example 3e),
ii. participles as complements of auxiliaries (see example 3c),
iii. the so-called *am* progressive (see example 3f),
iv. full NPs (see example 3g).

The constructions with the infinitive (i) and the participle (ii) represent two prototypically syntactic constructions as the verb from the N + V is part of a verbal syntagma. The NP context (iv) is most prototypically nominal, especially since we only used NPs with a determiner. More precisely, we only used NPs with definite determiners cliticised to a preposition (*beim* 'at the', *zum* 'to the', etc.). This decision was made in order to allow for a comparison of these full nominalisations and the so-called *am* progressive

[33] 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 small an influence on the estimation, 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 the practical difficulties of estimating a model on the raw 2,246,886 observations.

(iii). The progressive is formed with the copula/auxiliary *sein* 'to be', the variant of the preposition *an* with the cliticised definite article *am* 'at the' and the infinitive. While it developed out of a construction with a copula and a plain NP within a PP, and it is formally identical to cases with the normal NPs in (iv), it is often assumed to be a verbal construction expressing progressive meaning.[34] Including NPs in this specific form along with this emerging progressive construction allows us to assess whether the hypothesised verbal semantics of the progressive makes the construction more verbal, leading to a weaker tendency towards univerbations compared to regular NPs. To summarize, we expect N + V units in infinitives and participles (prototypically verbal) to have a weak tendency and N + V units in full NPs (prototypically nominal) to have a strong tendency towards univerbation in line with the argumentation from Section 2.2. We have no prediction for the progressive as we are unsure whether it has truly developed into a verbal construction.

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, ![Schäfer 2020]). We specified a generalised linear mixed model with the N + V unit variable as a random effect. The variable encoding the internal relation is nested inside the levels of the random effect, and its is therefore treated as a second-level fixed effect in a multilevel model. In R notation, the specification is shown in (16).[35]

(16) `Univerbation ~ (1|NVUnit) + Context + Relation`

The estimated parameters of the model are given in Table 2. Additionally, effect plots for *Context* and *Relation* are given in Figure 2.[36] As expected, the prototypically verbal contexts (infinitives and participles) are associated with a low probability of compound spelling (the infinitive is on the intercept, which is estimated at 1.054, and participles have a coefficient of 3.886). NPs and progressives as prototypically nominal contexts clearly favour compound spelling (coefficients of 4.907 and 1.344, respectively).

[34] The literature on the *am* progressive is rich. See Anthonissen, Wit & Mortelmans (2016) for an overview of the literature and a corpus-based assessment of its functions.

[35] See Appendix B for a precise specification in mathematical notation.

[36] 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.

	Estimate	CI low	CI high
.sig01	-4.685	2.002	2.222
(Intercept)	1.054	-4.895	-4.474
Context = Participle	3.886	0.975	1.133
Context = NP	4.907	3.815	3.959
Context = Progressive	1.344	4.801	5.015
Relation = Undetermined	3.085	0.866	1.822
Relation = Oblique	-4.685	2.764	3.407

Table 2: 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 N+V lemma: Intercept $= 4.442$, sd $= 2.108$. The intercepts model the fixed effects Relation = Argument. Nakagawa & Schielzeth's $R^2_m = 0.576$ and $R^2_c = 0.999$.

Both the coefficients and the effect plot (right panel in Figure 2) show a low probability of compound spelling when an argument relation holds between the verb and the noun (on the intercept), and a high probability when the relation is oblique (coefficient -4.685). The undetermined cases are in between the two clear-cut cases (coefficient 3.085).

Given the narrow confidence intervals and the high marginal measure of determination $R^2_m = 0.576$, we consider the hypotheses regarding fixed effects as well corroborated by the data, especially the effects of the context and the internal relation. The differences differences between specific N + V units already shown in Section 3.2 show up in the model as the residual variance in the random effects (in the form of the conditional modes).[37] The conditional modes are centred around a second-level intercept of 4.442 with a standard deviation of 2.108. The relatively high standard deviation is a sign that there is considerable variation across the individual N + V units. Furthermore, the conditional R^2_c is as high as 0.999. This is commonly interpreted as saying that the fixed effects and the idiosyncratic effect of concrete N + V units almost fully explain the variance in the data. A random selection of 20 conditional modes, which illustrates the relevance of lexical idiosyncrasies through obvious differences with mostly very narrow prediction intervals, is shown in Figure 3. The individual N + V unit thus plays a

[37] On a technical note, this is only the component of the variance which is not explained by the second-level effects.

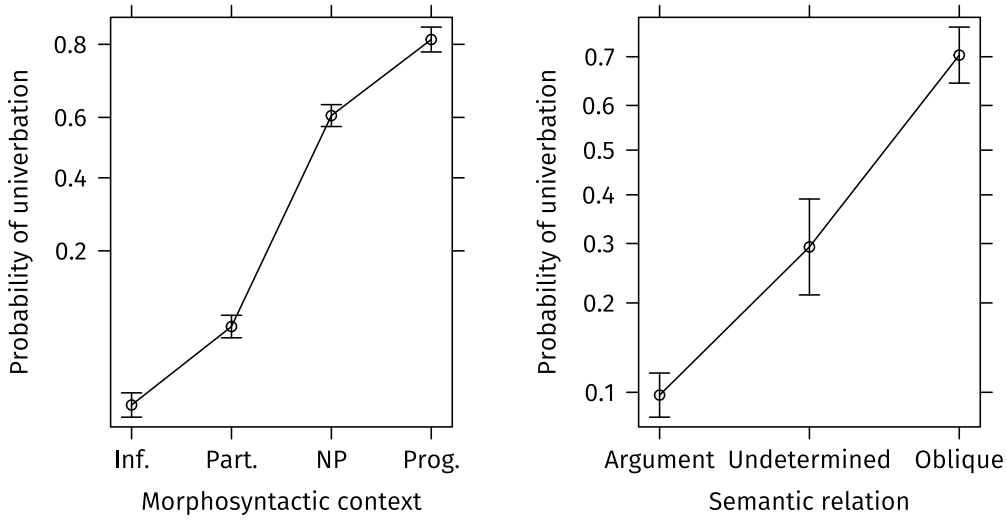

Figure 2: Effect plots for the regressor encoding the morphosyntactic context of the N+V unit and the regressor encoding the syntactic relation within the N+V unit in the GLMM modelling the corpus data.

major role in writers' tendency to univerbate N + V units, which conforms the results from Section 3.2.

4 Elicited production of noun-verb units

In this section, we corroborate the findings from Section 3 in a controlled experiment. We describe the rationale behind the experiment, the methods used, the design, and the group of participants in Section 4.1. Section 4.2 reports the results descriptively and in the form of a generalised linear mixed model.

4.1 Design and participants

The goal of the experiment was to replicate the findings from the corpus study in another empirical paradigm and to test whether writers' behaviour under controlled experimental conditions is similar to the behaviour of writers under uncontrolled circumstances as found in corpora. We used pre-recorded auditory stimuli in order to elicit spellings of given N + V units.

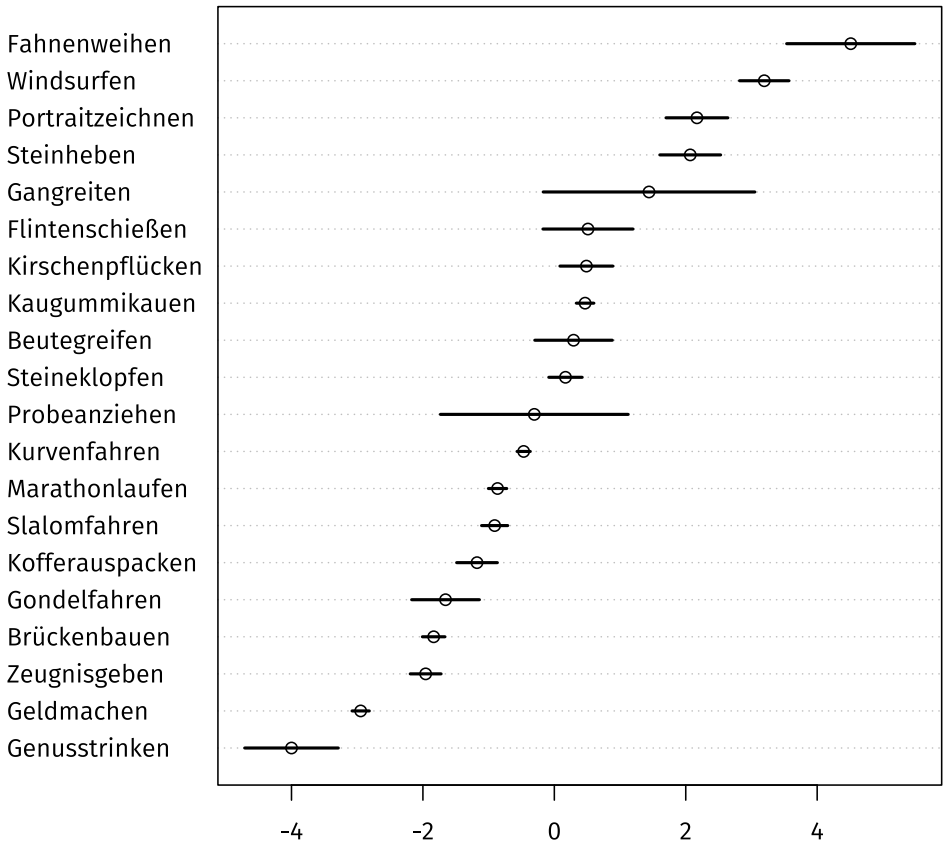

Figure 3: A random selection of conditional modes with 95% prediction intervals for the levels of the random effect in the GLMM modelling the corpus data.

The stimuli were chosen based on theoretically motivated criteria and the information about item-specific tendencies obtained from the exploratory part of the corpus study in Section 3.2. We constructed eight sentences instantiating the four morphosyntactic contexts described in Section 3.3 crossed with the two semantic relations.

Context	Relation	N+V unit	Attr. score
Infinitive	Argument	Platzmachen	-0.052
Infinitive	Oblique	Seilspringen	0.011
NP	Argument	Spaßhaben	-0.115
NP	Oblique	Bergsteigen	0.082
Participle	Argument	Mutmachen	-0.069
Participle	Oblique	Probehören	0.055
Progressive	Argument	Teetrinken	-0.037
Progressive	Oblique	Bogenschießen	0.087

Table 3: Items from the experiment, chosen by context and relation, with control for lexical attraction scores..

An overview of the item design is shown in Table 3, where each line represents the features of one of the eight items. In order to control for differences in lexical preferences, the concrete pairs of N + V units used in each context were chosen such that the contrast in lexical preference (see Section 3.2) for and against univerbation was as substantial as possible. As expected, units with an argument relation have negative attraction scores, and ones with an object relation have positive scores (see column "Attr. score" in Table 3). For each context, we selected pairs where the difference between the scores was larger than 0.05. Except for the infinitive context (difference 0.063), we managed to find pairs for which the difference is actually above 0.1 (NP: 0.197, Participle: 0.124, Progressive: 0.124).

The sentences were constructed in a way such that all N + V units were the predicate of a subordinate clause. This consistently ensured verb-last constituent order and avoided interfering verb-second effects, which are typical of independent sentences in German. The stimui with full glosses are given in Appendix A. Furthermore, we added 32 fillers, resulting in a total of forty sentences being read to the participants.[38] Of the forty sentences, twenty (including the target items) had to be written down by the participants. The order of the target items was randomised, but it was en-

[38] Of the fillers, six were actually target items from an unrelated experiment.

sured that there were at least three sentences in between two target stimuli. There were nine distractors in the form of yes-no questions related to random sentences previously heard by the participants.

In total, 61 participants took part in the experiment. All of them were 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 order of stimuli was different between the two groups.

4.2 Results

In this section, we report the parameter estimates of a GLMM modelling the behaviour of the participants in our experiment. The model specification in R notation is given in (17).[39] The coefficient estimates for the GLMM are reported in Table 4.

(17) `Univerbation` $\sim$ (`1`|`Participant`) + `Context` + `Relation`

	Estimate	CI low	CI high
(Intercept)	-10.316	-13.914	-7.839
Context = Participle	3.184	1.966	4.643
Context = NP	8.962	6.694	12.336
Context = Progressive	10.667	8.134	14.283
Relation = Oblique	6.951	5.054	10.078

Table 4: Coefficient table for the GLMM modelling the experiment data with 95% confidence intervals. Nakagawa & Schielzeth's $R^2_m = 0.836$ and $R^2_c = 0.910$. Random effect for participant: Intercept $= 2.717$, sd $= 1.648$ The intercept models the fixed effect Context = Infinitive as well as Relation = Argument.

There is some variation between writers as captured in the standard deviation of the conditional modes (1.648), but the small difference between the marginal R^2_m (0.836) and the conditional R^2_c (0.910) suggests that speaker variation does not explain much of the variance in the the data. This corroborates our assumption from Section 1 that the phenomenon is not about

[39] Appendix B provides the specification mathematical notation.

individuals mastering the norm to different degrees or having different preferences when it comes to univerbation. Instead, the major deciding factors are the ones predicted by our theoretical model.

There seems to be no evidence that the participle has a different effect than the infinitive (which is on the intercept) given the large confidence interval ([1.966..4.643]). On the other hand, progressives (10.667) and NPs (8.962) clearly have a much more positive effect on the probability of univerbation. Again, we do not see evidence for a any difference between NP and progressive contexts given the large and overlapping confidence intervals. The oblique relation favours univerbation as predicted (6.951) compared to the argument relation (which is modelled by the intercept), and despite a quite large confidence interval ([5.054..10.078]), the effect is clearly positive.

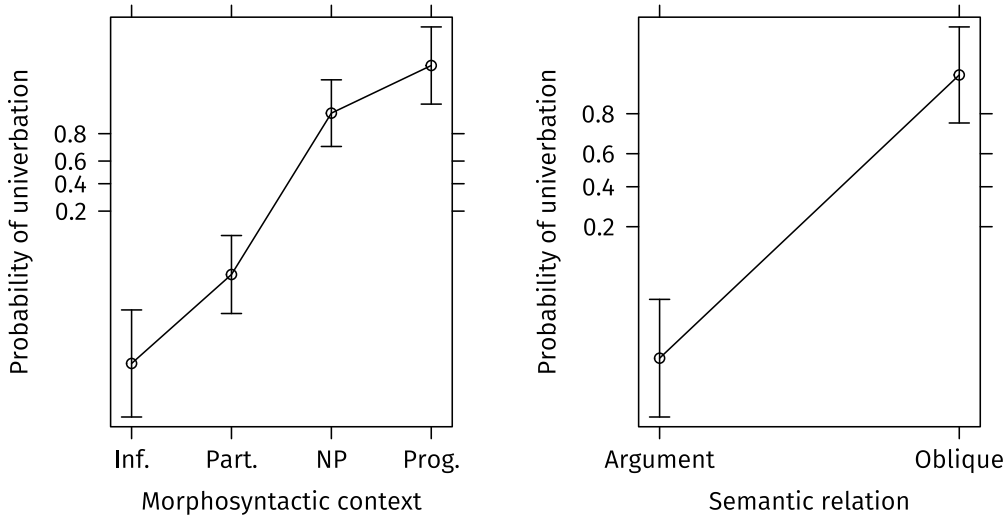

Figure 4: Effect plots for the regressor encoding the morphosyntactic context and the attraction strength as calculated from the corpus in the GLMM modelling the experimental data.

The effect plots in Figure 4 (left panel) provides a visual interpretation of the coefficient table. The prototypically verbal contexts are associated with low probabilities of univerbation, the two prototypically nominal ones with high probabilities of univerbation. Judging by the large and mostly overlapping confidence intervals, there is no support for assuming a substantial difference between participles and infinitives on the one hand and progres-

sives and NPs on the other hand. The two semantic relation are correlated with the probability of univerbation as expected (right panel of Figure 4).

In sum, the experiment supports our theoretically motivated hypotheses, and it corroborates the results from the corpus study. We proceed to a final analysis of the phenomenon at in the light of our findings in Section 5.

5 Explaining noun-verb univerbation

We have shown convincing evidence from corpora and controlled production experiments that the morphosyntactic context and the semantic relation are the crucial influencing factors on the graphemic univerbation of N + V units in German. Prototypically verbal contexts (infinitives and participles) disfavour univerbation, while prototypically nominal contexts (normal NPs and the so-called progressive, which contains a normal NP) favour univerbation. As we have argued, the nominal contexts enable the interpretation of the N + V unit as a proper compound (a morphological construction), while the verbal contexts are linked to a syntactic/phrasal interpretation. The difference in morphosyntactic status is mirrored in the different tendencies in writing. Furthermore, an argument relation between the V and the N within the N + V units disfavours univerbation because the N + V unit is closer to the regular syntactic construction than its counterpart with an oblique relation. For the oblique relation, a syntactic construction is barely accessible because it would normally require a preposition to mark the relation.

The fact that we could not find evidence for a difference in tendencies between the infinitive and participle speaks against a mixed verbal/nominal status of the participle in this specific construction, which does not preclude such a mixed status in other contexts.[40] The same goes for the *am*-progressive, which in our data behaves exactly like any other nominal construction. If it really is an emerging verbal syntagma (Anthonissen, Wit & Mortelmans 2016), this has no consequences for the NP status of the nominal element contained in it: it still behaves like a full NP in the context of the copula, at least in our data.

One aspect we have not yet discussed is the influence of the semantics of the verb. As a form of preliminary exploratory analysis, Figure 5 shows the distribution of N + V units with four selected head verbs.[41] The verb *machen* 'to make/do' clearly creates N + V units with weaker tendencies towards

[40] For participles as mixed categories, see Borik & Gehrke (2019).

[41] These plots are much like the one in Figure 1. However, the lower number of data points makes it infeasible to estimate a density curve. Instead, histograms were plotted.

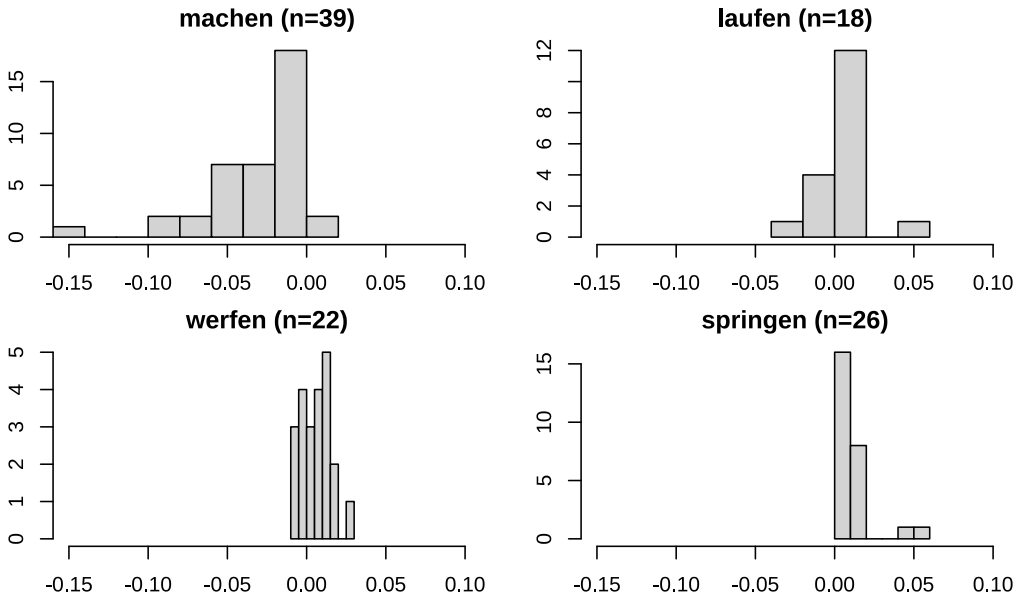

Figure 5: Distribution of attraction scores for N+V units with four different lexical verbs (*machen* 'to make/do', *laufen* 'to run/walk', *schießen* 'to shoot', *springen* 'to jump'); n is the number of N+V units with the respective V head in our corpus data.

univerbation, while *laufen* 'to run/walk' and *werfen* 'to throw' do not show a clear tendency, and *springen* 'to jump' has a tendency towards univerbation. This might be an indication that semantically weaker verbs like *machen* resist univerbation. However, the number of N + V units for each verb is too low to make any sound inferences, and an analysis in terms of (semantic) verb *classes* would be necessary. Given the difficulty of determining the appropriate verb classes, we save this for future work.

Another point that needs further examination in the future is the productivity of the N + V construction. Intuitively and from looking at the data, it appears that the units with an argument relation are formed much more productively compared to the ones with an oblique relation. If the ones with an oblique relation are formed less productively, they should have a tendency to be more strongly lexicalised, which might be a reason for their stronger tendency to univerbate. Related to the question of productivity, we might ask whether some of the N + V units are actually the result of back formation processes. For example, the verb *zwangsernähren* 'to force feed' is likely a back formation of the N + N compound *Zwangsernährung* 'force feeding' (with *Ernährung* 'feeding' being derived from *ernähren* 'to feed'), and it now appears as a N + V unit with the full array of finite and infinite verb forms. It is difficult to quantify the effect of such back formations on the present study. Clearly, further diachronic and synchronic research is required.

A final point we have not discussed prominently is the presence of so-called linking elements. They normally only appear between the nouns in N + N compounds, and while many of them look like plural markers of the first noun (*Tontaubenschießen* 'clay pigeon shooting' analysed as *Tontauben-schießen*, argument relation), others do not even look like inflectional forms of the first noun (*Leistungsschießen* '(high) performance shooting/competitive shooting' analysed as *Leistung-s-schießen*, oblique relation). Interestingly, these linking elements occur in some N + V units, but only in a minority. Table 5 shows which linking elements we found in our sample, and in how many of the units they occurred.

In principle, the linking element should be a very clear indicator of a fully nominal status and favour univerbation. However, they occur readily at least in infinite verb forms like *leistungsgeschossen* (participle), which means the linking element is adopted outside of its primary domain (nominal compounds), i. e., in verb forms. This might be yet another indication that we are dealing with back formation processes in some of these cases. In most of the cases with linking elements, a direct interpretation as a plural

Linking element	Plural-like	N+V units
(None)	No	617
-s	No	22
-(e)n	Yes	118
-e	Yes	44
-er	Yes	18

Table 5: Linking elements and the number of N+V units they occur in.

is also conceivable, and clearly, further research could shed light on this question.

In closing, we would like to posit that the kind of data that we find with respect to N + V units can only be explained satisfyingly within a usage-based probabilistic framework. It is clearly the function of the space in German writing to separate syntactic words, and hence univerbation is best explained as corresponding to the loss of syntactic independence and a crossing over to morphology. As the effect is clearly gradual (both diachronically and in the grammar of present-day writers), a probabilistic approach to grammar and the grammar-graphemics interface is required. The fact that we can name the influencing factors and provide a statistical model of their *systematic* (albeit non-categorical) strengths is very strong evidence for the alternation being encoded in cognitive grammars and not a processing effect or mere "performance". We are confident that future work will uncover many more probabilistic graphemics-grammar mappings like the one we discuss in this paper.

A 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 3.

(18) Lara trat zur Seite, um **Platz** zu **machen**.
Lara stepped to the side in order room to make
Lara stepped aside to make way.

(19) 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.

(20) 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.

(21) 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.

(22) 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.

(23) 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.

(24) 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.

(25) Simone hört ein Hörbuch, während sie am **Bogen schießen**
Simone listens an audiobook while she at the bow shoot
ist.
is
Simone listened to an audiobook while practicing archery.

B Full specifications of the models

In Section 3.3, the specification of the model was given in R notation as (16), repeated here as (26).

(26) `Univerbation ~ (1|NVUnit) + Context + Relation`

Another structurally identical generalized linear mixed model was specified in Section 4.2 as (17) and repeated here as (27).

(27) `Univerbation` ~ (1|`Participant`) + `Context` + `Relation`

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 R notation has to offer. Since the models are parallel in structure, we provide a precise specification for (26) and then point out the only major difference compared to (27).

Mathematically and thus more transparently, model (26) is given in (28). 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.

(28) $$Pr(Univ_i = 1) = logit^{-1}[\alpha_0 + \alpha_{NV_j[i]} + \vec{\beta}_{Cont} \cdot \vec{x}_{Cont_i}]$$

The probability of univerbation $Pr(Univ_i = 1)$ 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 found in observation i) and the dot product of the vector of dummy-coded binary value for the morphosyntactic context $\vec{x}_{Cont_i}$ and the vector of their corresponding regressors $\vec{\beta}_{Cont}$. Since it is a multilevel model, α_{NV} has its own linear model, which is given in (29).

(29) $$\alpha_{NV_j} = \gamma_j + \vec{\delta}_{Rel} \cdot \vec{x}_{Rel_j}$$

It is also assumed that (30) holds, i. e., that the random intercepts for individual N + V units are normally distributed.

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

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) and the dot product $\vec{\delta}_{Rel} \cdot \vec{x}_{Rel_j}$ of the vector of binary variables encoding the relation and the vector of their corresponding coefficients.

Since we do not have a nesting of the Relation predictor within a second-level effect in the case of (27), the Relation predictor becxomes a first-level effect, hence (31).

(31) $$Pr(Univ_i = 1) = logit^{-1}[\alpha_0 + \alpha_{Part_j[i]} + \vec{\beta}_{Cont} \cdot \vec{x}_{Cont_i} + \vec{\delta}_{Rel} \cdot \vec{x}_{Rel_j}]$$

Consequently, the second-level model is nothing more than the second-level intercept α_{Part_j}, which is also assumed to be a normally distributed random variable.

Acknowledgments

We thank Luise Rissmann for her help annotating and cleaning the corpus data as well as conducting most of the experiments.

References

Anthonissen, Lynn, Astrid De Wit & Tanja Mortelmans. (2016). Aspect meets modality: a semantic analysis of the German am-progressive. *Journal of Germanic Linguistics* 28(1). 1–30. http://dx.doi.org/10.1017/S1470542715000185.

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.

Berg, Kristian. (2016). Graphemic analysis and the spoken language bias. *Frontiers in Psychology* 7(388). 1–3. http://dx.doi.org/10.3389/fpsyg.2016.00388.

Borik, Olga & Berit Gehrke. (2019). Participles: form, use and meaning. *Glossa* 4(1: 109). 1–27. http://dx.doi.org/doi.org/10.5334/gjgl.1055.

Bredel, Ursula & Hartmut Günther. (2000). Quer über das Feld das Kopfadjunkt. Bemerkungen zu Peter Gallmanns Aufsatz Wortbegriff und Nomen-Verb-Verbindungen. *Zeitschrift für Sprachwissenschaft* 19(1). 103–110. http://dx.doi.org/10.1515/zfsw.2000.19.1.103.

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

Bybee, Joan L. & Clay Beckner. (2009). Usage-based theory. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 827–856. Oxford: Oxford University Press. http://dx.doi.org/10.1093/oxfordhb/9780199544004.013.0032.

Coulmas, Florian. (1996). Typology of writing systems. In Hartmut Günther, Otto Ludwig, Jürgen Baurmann, Florian Coulmas, Konrad Ehlich, Peter Eisenberg, Heinz W. Giese, Helmut Glück, Klaus B. Günther, Ulrich Knoop, Bernd Pompino-Marschall, Eckart Scheerer, Rüdiger Weingarten & Florian Coulmas (eds.), *Schrift und Schriftlichkeit: Ein interdisziplinäres Handbuch internationaler Forschung*, vol. 2, 1380–1387. Berlin & New York: De Gruyter Mouton.

Dąbrowska, Ewa. (2014). Words that go together: measuring individual differences in native speakers' knowledge of collocations. *The Mental Lexicon* 9(3). 401–418. http://dx.doi.org/10.1075/ml.9.3.02dab.

Dąbrowska, Ewa. (2016). Cognitive linguistics' seven deadly sins. *Cognitive Linguistics* 27(4). 479–491. http://dx.doi.org/10.1515/cog-2016-0059.

Dammel, Antje & Luise Kempf. (2018). Paradigmatic relationships in German action noun formation. *Journal of Word Formation* 2. 52–86. http://dx.doi.org/10.3726/zwjw.2018.02.02.

Divjak, Dagmar. (2016). Four challenges for usage-based linguistics. In Jocelyne Daems, Eline Zenner, Kris Heylen, Dirk Speelman & Hubert Cuyckens (eds.), *Change of paradigms – new paradoxes – recontextualizing language and linguistics*, 297–309. Berlin/Boston: De Gruyter Mouton. http://dx.doi.org/10.1515/9783110435597-017.

Divjak, Dagmar & Antti Arppe. (2013). Extracting prototypes from exemplars what can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2). 221–274. http://dx.doi.org/10.1515/cog-2013-0008.

Divjak, Dagmar, Ewa Dąbrowska & Antti Arppe. (2016). Machine meets man: evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27(1). 1–33. http://dx.doi.org/10.1515/cog-2015-0101.

Dobrić, Nikola. (2015). Three-factor prototypicality evaluation and the verb "look". *Language Sciences* 50. 1–11. http://dx.doi.org/10.1016/j.langsci.2014.12.005.

Dücker, Lisa & Renata Szczepaniak. (2017). "auffm teuffelß dantz haben sie auffr knotten korffen linen gedantzet". Die graphematische Markierung von Komposition in den Hexenverhörprotokollen aus dem 16./17. jh. *Jahrbuch für Germanistische Sprachgeschichte* 8(1). 30–51. http://dx.doi.org/10.1515/jbgsg-2017-0004.

Eisenberg, Peter. (2020). *Grundriss der deutschen Grammatik: Das Wort*. 5th edn. Stuttgart: Metzler. http://dx.doi.org/10.1007/978-3-476-05096-0.

Evert, Stefan. (2008). Corpora and collocations. In Anke Lüdeling & Maria Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, 1212–1248. Berlin: Mouton. http://dx.doi.org/10.1515/9783110213881.2.1212.

Fleischer, Wolfgang & Irmhild Barz. (2012). *Wortbildung der deutschen Gegenwartssprache*. Marianne Schröder (ed.). 4th edn. Berlin, Boston: De Gruyter. http://dx.doi.org/10.1515/9783110256659.

Ford, Marilyn & Joan Bresnan. (2013). Using convergent evidence from psycholinguistics and usage. In Manfred Krug & Julia Schlüter (eds.), *Research methods in language variation and change*, 295–312. Cambridge, MA: Cambridge University Press. http://dx.doi.org/10.1017/cbo9780511792519.020.

Fortmann, Christian. (2015). Verbal pseudo-compounds in German. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-formation: an international handbook of the languages of Europe*, vol. 1, 594–610. Berlin: De Gruyter Mouton. http://dx.doi.org/10.1515/9783110246254-036.

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. http://dx.doi.org/10.18637/jss.v087.i09.

Fuhrhop, Nanna. (2007). *Zwischen Wort und Syntagma. Zur grammatischen Fundierung der Getrennt- und Zusammenschreibung*. Tübingen: Niemeyer. http://dx.doi.org/10.1515/9783110936544.

Gaeta, Livio. (2010). Synthetic compounds: with special reference to German. In Sergio Scalise & Irene Vogel (eds.), *Cross-disciplinary issues in compounding*, 219–2366. Amsterdam: Benjamins. http://dx.doi.org/10.1075/cilt.311.17gae.

Gaeta, Livio & Barbara Schlücker (eds.). (2012). *Das Deutsche als kompositionsfreudige Sprache: strukturelle Eigenschaften und systembezogene Aspekte*. Berlin: De Gruyter. http://dx.doi.org/10.1515/9783110278439.

Gaeta, Livio & Amir Zeldes. (2017). Between VP and NN: on the constructional types of German -er compounds. *Constructions and Frames* 9(1). 1–40. http://dx.doi.org/doi10.1075/cf.9.1.01gae.

Gallmann, Peter. (1999). Wortbegriff und Nomen-Verb-Verbindungen. *Zeitschrift für Sprachwissenschaft* 18(2). 269–304. http://dx.doi.org/10.1515/zfsw.1999.18.2.269.

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

Gilquin, Gaëtanelle. (2006). The place of prototypicality in corpus linguistics: causation in the hot seat. In Stefan Th. Gries & Anatol Stefanowitsch (eds.), *Corpora in cognitive linguistics: corpus-based approaches to syntax and lexis*, 159–191. De Gruyter Mouton.

Gries, Stefan Th. (2003). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1. 1–27. http://dx.doi.org/10.1075/arcl.1.02gri.

Gries, Stefan Th. (2015). More (old and new) misunderstandings of collostructional analysis: on Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536. http://dx.doi.org/10.1515/cog-2014-0092.

Günther, Hartmut. (1997). Zur grammatischen Basis der Getrennt-/Zusammenschreibung im Deutschen. In Christa Dürscheid (ed.), *Sprache im Fokus: Festschrift für Heinz Vater zum 65. Geburtstag*, 3–16. Tübingen: Niemeyer.

Haugen, Jason D. (2015). Incorporation. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen & Franz Rainer (eds.), *Word-formation: an international handbook of the languages of Europe*, vol. 1, 413–434. Berlin: De Gruyter Mouton. http://dx.doi.org/10.1515/9783110246254-024.

Hentschel, Elke & Harald Weydt. (2003). *Handbuch der deutschen Grammatik*. 3rd edn. Berlin, Boston: De Gruyter. http://dx.doi.org/10.1515/9783110312973.

Hoberg, Ursula. (1981). *Die Wortstellung in der geschriebenen deutschen Gegenwartssprache*. München: Hueber.

Hüning, Matthias. (2010). Adjective + Noun constructions between syntax and word formation in Dutch and German. In Alexander Onysko & Sascha Michel (eds.), *Cognitive perspectives on word formation*, 195–216. Berlin, New York: De Gruyter Mouton. http://dx.doi.org/doi.org/10.1515/9783110223606.195.

Jacobs, Joachim. (2005). *Spatien. zum System der Getrennt- und Zusammenschreibung im heutigen Deutsch*. Berlin: De Gruyter. http://dx.doi.org/10.1515/9783110919295.

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

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.
Klos, Verena. (2011). *Komposition und Kompositionalität. Möglichkeiten und Grenzen der semantischen Dekodierung von Substantivkomposita*. Berlin, New York: De Gruyter. http://dx.doi.org/10.1515/9783110258875.
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. http://dx.doi.org/10.1515/cog-2015-0053.
Lehmann, Christian. (2020). Univerbation. *Folia Linguistica Historica* 42(1). 205–252. http://dx.doi.org/10.1515/flih-2020-0007.
Mithun, Marianne. (1984). The evolution of noun incorporation. *Language* 60(4). 847–894. http://dx.doi.org/10.2307/413800.
Mithun, Marianne. (2000). Incorporation. In Geert Booij, Christian Lehmann & Joachim Mugdan (eds.), *Morphologie: ein internationales Handbuch zur Flexion und Wortbildung, vol. 1*, 916–928. Berlin: De Gruyter Mouton. http://dx.doi.org/10.1515/9783110111286.1.12.916.
Morcinek, Bettina. (2012). Getrennt- und Zusammenschreibung: Wie aus syntaktischen Strukturen komplexe Verben wurden. In Livio Gaeta & Barbara Schlücker (eds.), *Das Deutsche als kompositionsfreudige Sprache. Strukturelle Eigenschaften und systembezogene Aspekte*, 83–100. Berlin: De Gruyter. http://dx.doi.org/10.1515/9783110278439.83.
Murphy, Gregory. (2002). *The big book of concepts*. Cambridge: MIT Press.
Muthmann, Gustav. (1988). *Rückläufiges deutsches Wörterbuch. Handbuch der Wortausgänge im Deutschen mit Beachtung der Wort- und Lautstruktur*. Tübingen: Niemeyer. http://dx.doi.org/10.1515/9783110920666.
Newman, John. (2011). Corpora and cognitive linguistics. *Revista Brasileira de Linguística Aplicada* 11(2). 521–559. http://dx.doi.org/10.1590/S1984-63982011000200010.
Nübling, Damaris, Antje Dammel, Janet Duke & Renata Szczepaniak. (2017). *Historische Sprachwissenschaft des Deutschen. Eine Einführung in die Prinzipien des Sprachwandels*. Tübingen: Narr.
Pankratz, Elizabeth & Bob Van Tiel. (2021). The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition* 13(4). 562–594. http://dx.doi.org/10.1017/langcog.2021.13.
Pauly, Dennis Nikolas & Guido Nottbusch. (2020). The influence of the German capitalization rules on reading. *Frontiers in Communication* 5(15). 1–15. http://dx.doi.org/10.3389/fcomm.2020.00015.
Primus, Beatrice. (2010). Strukturelle Grundlagen des deutschen Schriftsystems. In Ursula Bredel, Astrid Müller & Gabriele Hinney (eds.), *Schrift-*

system und Schrifterwerb, 9–45. Berlin, New York: De Gruyter. http://dx.doi.org/10.1515/9783110232257.9.

Rosch, Eleanor. (1973). Natural categories. *Cognitive Psychology* 4(3). 328–350. http://dx.doi.org/10.1016/0010-0285(73)90017-0.

Rosch, Eleanor. (1978). Principles of categorization. In Eleanor Rosch & Barbara B. Lloyd (eds.), *Cognition and categorization*, 27–48. Hillsdale: Erlbaum. http://dx.doi.org/10.1016/b978-1-4832-1446-7.50028-5.

Schäfer, Roland & Ulrike Sayatz. (2014). Die Kurzformen des Indefinitartikels im Deutschen. *Zeitschrift für Sprachwissenschaft* 33(2). 215–250. http://dx.doi.org/10.1515/zfs-2014-0008.

Schäfer, Roland & Ulrike Sayatz. (2016). Punctuation and syntactic structure in "obwohl" and "weil" clauses in nonstandard written German. *Written Language and Literacy* 19(2). 212–245. http://dx.doi.org/10.1075/wll.19.2.04sch.

Schäfer, Roland. (2018). Abstractions and exemplars: the measure noun phrase alternation in German. *Cognitive Linguistics* 29(4). 729–771. http://dx.doi.org/10.1515/cog-2017-0050.

Schäfer, Roland. (2019). Prototype-driven alternations: the case of German weak nouns. *Corpus Linguistics and Linguistic Theory* 15(2). 383–418. http://dx.doi.org/10.1515/cllt-2015-0051.

Schäfer, Roland. (2020). Generalized linear mixed models. In Stefan Gries & Magali Paquot (eds.), *A practical handbook of corpus linguistics*, 535–561. Berlin, Heidelberg: Springer. http://dx.doi.org/10.1007/978-3-030-46216-1_22.

Schäfer, Roland & Felix Bildhauer. (2012). Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC '12)*, 486–493. Istanbul: 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.

Scherer, Carmen. (2012). Vom Reisezentrum zum Reise Zentrum – Variation in der Schreibung von N + N-Komposita. In Livio Gaeta & Barbara Schlücker (eds.), 57–81. Berlin: De Gruyter. http://dx.doi.org/10.1515/9783110278439.57.

Schlücker, Barbara. (2012). Die deutsche Kompositionsfreudigkeit: Übersicht und Einführung. In Livio Gaeta & Barbara Schlücker (eds.), 1–25. Berlin: De Gruyter. http://dx.doi.org/10.1515/9783110278439.

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.

Simunic, Roman Nino. (2018). *Datenakquisition und Datenanalyse von Nomen-Adjektiv-Komposita*. Bochum: Ruhr-Universität Bochum PhD thesis.

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.

Stefanowitsch, Anatol & Stefan Th. Gries. (2009). Corpora and grammar. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics: an international handbook*, vol. 2, 933–952. Berlin: De Gruyter Mouton. http://dx.doi.org/10.1515/9783110213881.2.933.

Stumpf, Sören. (2015). *Formelhafte (Ir-)Regularitäten. Korpuslinguistische Befunde und sprachtheoretische Überlegungen*. Frankfurt am Main: Peter Lang.

Sutcliffe, John P. (1993). Concepts, class, and category in the tradition of Aristotle. In Iven Van Mechelen, James A. Hampton, Ryszard S. Michalski & Peter Theuns (eds.), *Categories and concepts: theoretical views and inductive data analysis*, 35–65. London: Academic Press.

Szczepaniak, Renata. (2009). *Grammatikalisierung im Deutschen. Eine Einführung*. Tübingen: Narr.

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

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

Tomasello, Michael. (2003). *Constructing a language: a usage-based theory of language acquisition*. Harvard: Harvard University Press.

Vogel, Petra Maria. (2000). Nominal abstracts and gender in Modern German: a "qualitative" approach towards the function of gender. In Barbara Unterbeck (ed.), *Gender in grammar and cognition*, 461–493. Berlin, New York: De Gruyter Mouton. http://dx.doi.org/10.1515/9783110802603.461.

Werner, Martina, Veronika Mattes & Katharina Korecky-Kröll. (2020). The development of synthetic compounds in German: relating diachrony with LI acquisition. *Word Structure* 13(2). 166–188.

Wurzel, Wolfgang Ullrich. (1994). Inkorporierung und "Wortigkeit" im Deutschen. In Wolfgang U. Dressler (ed.), *Natural morphology: perspectives for the nineties*, 109–125. Wien: Unipress.

Wurzel, Wolfgang Ullrich. (1998). On the development of incorporating structures in German. In Richard M. Hogg & Linda van Bergen (eds.), *Historical linguistics 1995*, 331–344. Amsterdam, Philadelphia: Benjamins. http://dx.doi.org/10.1075/cilt.162.24wur.

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.