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 …

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

1 Introduction

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) only the input and general cognitive constraints determine the grammar. Since the input is always rife with variation and often non-discrete, 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. While the phono-component comprises regularities of how grammar is encoded in speech sounds, graphemics comprises similar regularities of 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,

REFERENCE?

where 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. For writing systems like German, the mappings to be learned include sounds to letters, parts of speech to spellings, syntactic categories to spaces and punctuation marks, etc.[1]

REFERENCE?

REFERENCE?

REFERENCE FOR FN?

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; !Schäfer 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 competent 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 corpus data as evidence and assuming that they directly mirror the 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 performed by a majority of the learners, and 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 inasmuch as 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.[2] 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 sim-

[1] 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.

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

ilarity to a protoypical 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 stronger a unit matches the protoype, the more likely it is to be realised as mapped.

One caveat that is specific to graphemics needs to be dealt with before we proceed to the description of the concrete phenomena. The acquisition of the writing system involves explicit instruction and is thus more strongly superimposed by prescriptive norms. However, we expect writers to learn grammar-graphemics mappings first and foremost from their realisations in the input, especially whenever the norm is unspecific or unclear, a situation which provides ideal test cases for our view on graphemics. Variation or alternation in the written input shapes 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.[3]

The alternation we are going to explore affects units containing a verb and a noun, and these units alternate between a syntactic manifestation (where the noun combines with the verb via a syntactic relation) 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-sakting/ice-skates.

In this construction, there is a noun N occurring in strictly bare form, which is either an argument (normally in the accusative case) as in (1a) or an adjunct of the verb V as in (1b), which would normally take the form of a prepositional phrase.[4] We use the terms ‘argument relation’ and ‘oblique

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

[4] Whereas singular indefinite mass nouns typically occur without an article in German (Vogel 2000: 471), this is the only frequent construction in German wherein bare count nouns

relation' to refer to the semantic relation between the noun and the verb following Gaeta & Zeldes (2017: 20). Nouns with oblique status 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 (2).[5]

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

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

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

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

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.

[5] Further spelling variants for (2c) through (2g) will be discussed immediately below.

Such N + V units occur flexibly in all types of syntactic contexts: with finite verbs in verb-second order (2a), with finite verbs in verb-last order (2b), in the analytical perfect where the lexical verb takes the form of a participle (2c), in infinitives with the particle *zu* (2d), in a progressive-like construction with the preposition *an* fusioned with the dative singular article *dem* to *am* where the infinitive is potentially nominalised (2e), and in regular prepositional phrases (2f). In (2g), 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.[6] The spelling as two words for (2e) and (2f) is not accepted by all native speakers, a fact to which will return throughout the paper.

In the examples (2c) through (2g), 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 (2) and spellings as one word shown in (3).

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

In (3e) and (3f), additional variation is introduced in the form of upper-case and lower-case initials.[7] The compound with the nominalised infinitive in (3g) is fine if spelled as one word.

We call cases where a multi-stem unit is spelled as two words such as in (2) 'disjunct spellings' and cases where a unit is spelled as one word as in (3) 'compound spelling'. We see that N + V units potentially undergo graphemic *univerbation* in the form of compound spelling. Lehmann (2021: 2) calls univerbation "the union of two syntagmatically adjacent word forms in one". We follow this terminology and assume univerbation 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 compound spelling. Historically and—as we're going to show especially in Section 4—individually, univerbation is a gradual process, and it is thus a probabilistic phenomenon. However, univerbation per se is not

[6] 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).

[7] In German, all nouns are capitalised anywhere in a sentence (Pauly & Nottbusch 2020: 1).

necessarily the result of a regular grammatical pattern.[8] 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. This double nature and the resulting alternation is interpreted as favouring probabilistic competence models over deterministic competence models. In Section 2, we lay the theoretical and descriptive foundations. Section **??** introduces probabilistic grammar and how graphemic evidence can be interpreted in probabilistic grammar. In Section 2.1, the nature and status of spaces and their loss (univerbation) are discussed briefly, and Section 2.2 provides a detailed account of N + V units in German. We finish Section 2.2 by summing up our hypotheses before presenting a large-scale corpus study and an elicitation experiment in Sections 3 and 4, both representing tests our particular hypotheses about N + V units and the overarching hypothesis regarding the probabilistic nature of grammar. We conclude with a summary, further interpretation and discussion in Section 5.

2 Theoretical background

2.1 Spaces, words, and univerbation

In this paper, we use graphemic evidence—both from corpora and from controlled experiments—and argue that it allows us to draw conclusions about writers' cognitive grammars. More specifically, we assume that compound spellings of N + V units indicate that writers conceive 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.

[8] 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 (2021: 4) 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 phonological fusion.

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.[9] 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, compound spelling of syntactic words comprising more than one stem was also established in the history of German writing, 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*.[10] For such cases, Lehmann (2021: 2) 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 conventionalised as a single lexical word or—in other words—lexicalised (Lehmann 2021: 7). Until such a diachronic process is complete and one of the spellings has become clearly dominant, conventionalisation does not provide a very strong input to writers, and they alternate between a syntagmatic and a morphological realisation. For many of these

[9] 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).

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

constructions, this is the case both in non-standard as well as standard written German, albeit to different degrees.[11]

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.[12] 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 token 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 by individual N + V units and their syntactic contexts more or less well, which leads to either compound or disjunct spelling being the preferred realisation. The hypotheses put forward here are then tested in Sections 3 and 4.

In order to achieve this end, we need to shed some light on the productive N + N compound construction in Section 2.2.1 before turning to N + V

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

[12] 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 *Auto fahren* was supposed to be spelled as two words exclusively (Primus 2010: 32, Eisenberg 2020: 356).

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 undergo graphemic univerbation (i. e., a downgrading of a syntactic to a morphological construction in the sense of Lehmann 2021: 2) systematically, it must resemble one or more established morphological constructions closely enough to be classified as an instance of such constructions itself.[13] 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.[14] German N + N compounds instantiate a proper morphological construction and are therefore inseparable. 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 (4) and (5) for *Haustür* and *Haustürschlüssel*, the latter being recursively formed from the former.[15]

(4) Haus.tür
house.door
front door

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

[13] 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.

[14] 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 don't discuss it here. See Simunic (2018: 136) on the productivity of N + A compounds.

[15] If necessary, we present compound spelling with a minimal anaylsis of the morphological structure. Affixes are separated from stems by hyphens, an lexical stemsare separated from each other by a period. Within compounds containing more than two stems, structure is using using square brackets as in examples (8) and (9) below.

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).[16] The historic development of the stable N + N compound construction was furthered during the Early High German period (approximately from the 14th to the 17th century AD) by a syntactic change.[17] The dominant pattern of noun–noun attribution had been a prenominal genitive as in the now obsolete (6), which swiftly changed to a postnominal genitive as in (7).

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

(7) die Tür des Hauses
the door the$_{\text{Gen}}$ 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 (4) (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 are just concatenated as bare stems in most cases, but there are also so-called linking elements, which are sometimes positioned in between the stems.[18] Diachronically, linking elements stem 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,

[16] 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').

[17] Nübling et al. (2017: 130) finds 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.

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

and how this leads to them alternating between a syntactic and a morphological construction.

2.2.2 N+V units as reluctant compounds

In this section, we argue that N + V units are *reluctant compounds*. While in principle the morphological N + V construction (as a kind of compound written as one word) has existed for centuries, we show why and how it remains in competition with a syntactic construction. At the same time, we argue why—at least under the right circumstances—morphological compounding (and consequently the spelling as one word) become the preferred realisation.

Most likely, full compounding of N + V units requires conversion of the verbal head to a noun 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.[19] A major difference compared 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 syntactic material in between the noun and the verb in some contexts, namely the infinitival particle *zu*. Furthermore, the noun and the verb can be reordered in verb-second order, where N + V units resemble particle verbs (Fortmann 2015: 603), see (**??**). This fact alone means that N + V units do not fit the compounding prototype well. This likely introduces great resistance in speakers to classify them as compounds and consequently 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 + N compounds (contrary to claims by Fuhrhop 2007: 54) as in (8).[20] 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 (9). Native speakers will readily acknowledge that such constructions are outright absurd.

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

[20] The examples in (8) 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.

(8) 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

(9) * 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 as well as both unspecified and often unspec ific (see Section 2.2.1), there are only two possible relations within N + V units, and these relations are determined by—and above all decodable through—the distributional properties of the verb (including its argument structure). It's either an object relation or an adjunct relation where all distributional restrictions apply that would apply in a syntactic realisation of the same verb. As a consequence, there is always a syntactic paraphrase for N + V units with an adjunct relation where the noun occurs in a prepositional phrase which is an adjunct to the verb.[21] See (10) for an example.

(10) 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.

[21] 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.

The relation is decodable except in rare cases which underwent full lexicalisation a long time ago such that the meanings of the lexemes or their distributions have changed significantly. However, the decodable relations (direct object or prepositional adjunct) are prototypically realised syntactically as German is a language with a very weak (if any) tendency towards noun incorporation (see below in this section). The arguments and adjuncts of a verb are usually realised as syntactic dependants of the verb. Even if the verb is nominalised, direct objects are realised as genitives in the noun phrase, and adjuncts don't change their form when occurring as dependants of nominalised verbs.

The fact that the relation can be decoded for almost all N + V units means that the morphological construction marked by graphemic univerbation almost always 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 parallel syntactic construction for N + N compounds (prenominal genitives) disappeared within a relatively short period of time, the ambiguity between syntax and morphology of N + V units remains intact. This is true although univerbation of N + V units with an object relation dates back to Middle High German (*lobpreisen* 'praise', literally 'to praise compliment') and even Old High German (*hals-werfōn* 'turning around', literally 'to turn neck'), see Wurzel (1994: 517), Wurzel (1998: 334). For N + V units with an adjunct relation, Morcinek (2012: 89) notices that dictionaries from between 1750 and 1993 AD list novel N + V units with an adjunct relation with increasing frequency. For centuries or 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 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*.

We still have to show why and under which conditions we assume true compounding and (including potential V-to-N conversion of the head and subseqeunt graphemic univerbation) to be preferred. As mentioned in Section 1, the morphological construction for 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 postulated above that object and prepositional adjunct relations are prototypically realised syntactically. 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”.[22] 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 the semantics of which are determined by the previous syntactic relation between the noun and the verb. Typically, the noun looses 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 (11), 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.

(11) 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 (12).

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

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. Hence, N + V units have a tendency to form proper compounds and subsequently undergo univerbation despite the factors mentioned above that pull them towards a syntactic construction.

In the next section, we will summarise the factors that influence the tendency of N + V units to incorporate and undergo univerbation. We also formulate testable hypotheses for the empirical work to be reported in Sections 3 and 4.

[22] From Mithun’s types of noun incorporation, German N + V units clearly represent type 1 *lexical compounding*. Since nothing could be gained from it, we do not discuss the literature on the typological classification of incorporation further.

2.2.3 Conclusions for the empirical studies

In the previous sections, we have laid out a theory of the factors leading to the univerbation of N + V units. In this section, we derive some 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.[23] 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. As argued, N + V units with an oblique relation stand in weaker competition with a syntactic realisation compared with those that have an object relation. N + V units with an oblique relation would need more explicit marking with a preposition in the alternative unambiguously syntactic realisation. Also, for N + V units with an argument relation, there is the productive and functionally similar type of government $N_1 + N_2$ compound where N_2 is usually derived from a verb. However, there is no similar morphological pattern for units with an adjunct relation. Hence, we expect a stronger tendency towards univerbation with oblique relations because they have a more accessible interpretation as a morphological unit (a compound).

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 protoypicality effects described here, individual units might have progressed further than others on the lexicalisation path. Furthermore, when the compositional meaning of individual N + V units becomes less accessible, univerbation might be favoured due to a facilitated emergence of a holistic (semantically incorporated) conceptual semantics of the unit. Also, units with semantically weak or generic verbs like *haben* (‘to have’), *machen* (‘make’), and *fahren* (‘drive’) are ex-

[23] Notice that when the V head is finite, the N and the V are always realised discontontinously, and univerbation is not an option. The most prototypical verbal realisation is thus outside the scope of this study.

pected to undergo univerbation more easily, because the verbs only denote a specific concept together with the noun. As 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 presence of a linking element cannot be conceived of as a causal influencing factor, but rather an additional indicator of compound status. Since the linking *-s* is not paradigmatic in N + V units realised syntactically, its presence is expected in cases where the unit is used as a true N + N compound.

Finally, individual speakers should be expected to have individual tendencies due to the variance in their input and 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 only gradually. Even in cases where all factors favour a realisation with univerbation, writers might spell it without univerbation and vice versa. In usage data, such cases are just expected to be rare if the hypotheses put forward here correctly describe reality.

As a preliminary step, the basis for any empirical look at N + V units and their spelling has to be a data-driven assessment of which units exist, how strongly they alternate, and what their item-specific tendencies are, i. e., how clearly they tend to be spelled as one word or two words. Therefore, Section 3 begins with such an assessment.

3 Analysing the usage of noun-verb units

In this section, we use 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 incorporate and undergo univerbation in written language usage. The tendencies calculated here will also be used as a control variable in the experiment reported in Section 4. 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 (for example with modal verbs).

Clearly, a large corpus with rich morphological and morphosyntactic annotations containing texts written in a broad variety of registers and styles (including ones written under low normative pressure) was required. We chose the DECOW16B corpus (Schäfer & Bildhauer 2012) because it fulfils all the aforementioned criteria.[24] 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 searches of roots within nominal compounds. For example, we could query compounds with a deverbal head such as *Zeitnehmen* (‘time taking’). Furthermore, the interface offered by the creators of the COW corpora allows for automated queries controlled by Python scripts using the open-source SeaCOW interface.[25] The scripts we used to make the queries are released on a curated open-data server along with all data as well as the LaTeX, knitr, and R scripts created in the writing of this paper.[26]

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

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

[25] https://github.com/rsling/seacow

[26] The DOI of the data set will be revealed in the accepted version of this paper.

[27] See the scripts available under the abovementioned DOI for concrete queries and further details.

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 V + N unit and used these to query all possible compound and separate spellings (including variance in capitalisation) of each of the 819 N + V unit types. In total, 28,665 queries were executed to create the final data set used here, a number which clearly demonstrates the necessity of script-based corpus access in data-driven methods. The queries were matched by 958,118 compound spellings and 1,288,768 separate spellings, which results in a total sample size of 2,246,886 tokens.

For each N + V unit in the sample, the following variables were annotated automatically: (i) the verb lemma, (ii) the noun lemma, (iii) whether a linking element is used in the use as a full noun, (iv) the overall frequency in the corpus. Additionally, we manually coded all 819 N + V units for the relation holding between the verb and the noun. The codes used in clear-cut cases were *Object* (441 units) and *Adjunct* (286 units). For 92 units, both relations were conceivable, and those cases were coded as *Undetermined*. This class is illustrated by *Daumenlutschen* ("thumb sucking"), which could correspond to the paraphrase either in (13a) or in (13b).

(13) 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. Second, in Section 3.3 we report a full statistical model of the alternation. Finally, the association strengths calculated in Section 3.2 were also used as a covariate in the analysis of the experiment reported in Section 4.

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 collocational

[28] Notice that three highly frequent N + V units were excluded because they could be considered outliers, having an overly strong tendency to be used in compound spelling. They are *Teilnehmen* 'to take part', *Maßnehmen* 'take measure', and *Teilhaben* 'have part' (meaning 'to participate'.

analysis (Evert 2008 for an overview) and stems from under the umbrella of collostructional analysis (Stefanowitsch & Gries 2003). More specifically, the method is called *collexem 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 number 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 proportion 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 its 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

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

[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).

the 819 N + V units. Their distribution is plotted in the form of a density estimate in Figure 1.[31]

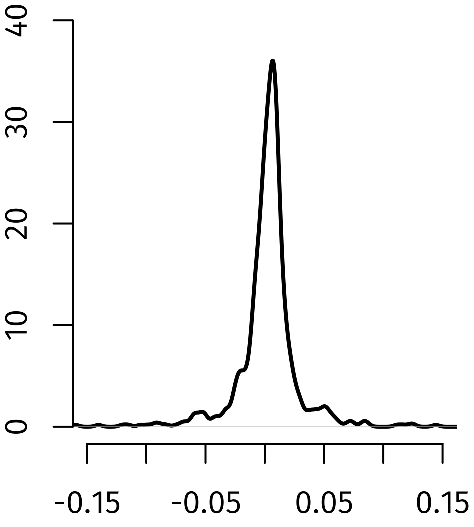

Figure 1: Density estimate of the distribution of the 819 association scores (across all morphosyntactic conditions).

For the selection of the target stimuli used in the experiment, the these association strengths are important, because they encode the effect of the individual N + V units, independently of the context. Context effects were controlled independently in the experiment. To illustrate how the data analysis allows for a selection of N + V units based on their affinity towards univerbation, we show the top ten units with the highest negative and highest positive association in Table 2. The tables illustrate that units with the strongest tendencies against univerbation are predominantly ones with an object relation. The ones which most strongly favour univerbation are mostly ones with an adjunct relation or an ambiguous relation. The ten items with the least clear tendency in either direction are shown in Table 3. They mostly come with an object relation.

Among the units with an object relation, it is difficult to tell based on native-speaker intuition why the ones in Table 3 should have no preference and the ones in Table 2 (right panel) should resist univerbation. This goes to show that, while we can model the tendencies to some extent using linguis-

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

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

Table 2: Top ten V+N units with a strong tendency for univerbation (left panel) and top ten V+N units with a strong tendency against univerbation (right panel).

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

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

tic features, there are obvious item-specific effects which should be taken seriously from a theoretical perspective, and which must be accounted for in behavioural experiments. We turn to such an experiment in Section 4. However, we first report a full statistical model of the influencing factors derived from the corpus data in Section 3.3.

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 concice interpretation and allows for finer-grained data analysis. Also, it has long been accepted that using several methods strengthens the analysis when the results converge (e. g., Arppe & Järvikivi 2007).

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, the presence of absence of a linking element in the nominal compound (as a marker of a stronger lexicalisation) 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

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

[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

(or observation-level) fixed effect regressor is the morphosyntactic context (a four-way categorical variable). As there is a huge number of 819 N + V units, the lexical indicator variable for the individual N + V unit should not be used as a fixed effect (Gelman & Hill 2006: 244–247, !Schäfer 2020). We specified a generalised linear mixed model with the N + V unit variable as a random effect. The variables encoding the internal relation and the presence/absence of a linking element are nested inside the levels of the random effect, and they are therefore treated as second-level fixed effects in a multilevel model. In R notation, the specification is shown in (14).[34]

(14) $\texttt{Univerbation} \sim (\texttt{1}|\texttt{NVUnit}) + \texttt{Context} + \texttt{Relation} + \texttt{Link}$

	Estimate	CI low	CI high
(Intercept)	-4.787	2.000	2.218
Context = Participle	1.054	-5.020	-4.555
Context = NP	3.886	0.976	1.133
Context = Progressive	4.907	3.815	3.959
Relation = Undetermined	1.339	4.801	5.015
Relation = Adjunct	3.132	0.862	1.816
Link = Yes	0.361	2.808	3.456

Table 4: Coefficient table for the binomial GLMM modelling the corpus data with 95% profile likelihood ratio confidence intervals. The horizontal line separates first-level and second-level effects. Weighting was used to account for the bias in models on proportion data. Random effect for V+N lemma: $\text{Intercept} = 4.430$, $\text{sd} = 2.105$. The intercepts model the fixed effects Relation = Object and Link = No. Nakagawa & Schielzeth's $R^2_m = 0.577$ and $R^2_c = 0.999$.

The estimated parameters of the model are given in Table 4. Additionally, effect plots for *Context* and *Relation* are given in Figure 2.[35] As

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.

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

[35] 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

expected, the prototypically verbal contexts (infinitives and participles in analytic verb forms) are associated with a low probability of compound spelling (the infinitive is on the intercept, which is estimated at −4.787, and participles have a coefficient of 1.054). NPs and progressives as prototypically nominal contexts clearly favour compound spelling (coefficients of 3.886 and 4.907, respectively). Both the coefficients and the effect plot (right panel in Figure 2) show a low probability of compound spelling when the relation between the verb and the noun (on the intercept) is that of an object, and a high probability when the relation is that of an adjunct (coefficient 3.132). The undetermined cases are in between the two clear-cut cases (coefficient 1.339). The presence of a linking element in fully nominalised compounds favours compound spelling only slightly (coefficient 0.361).

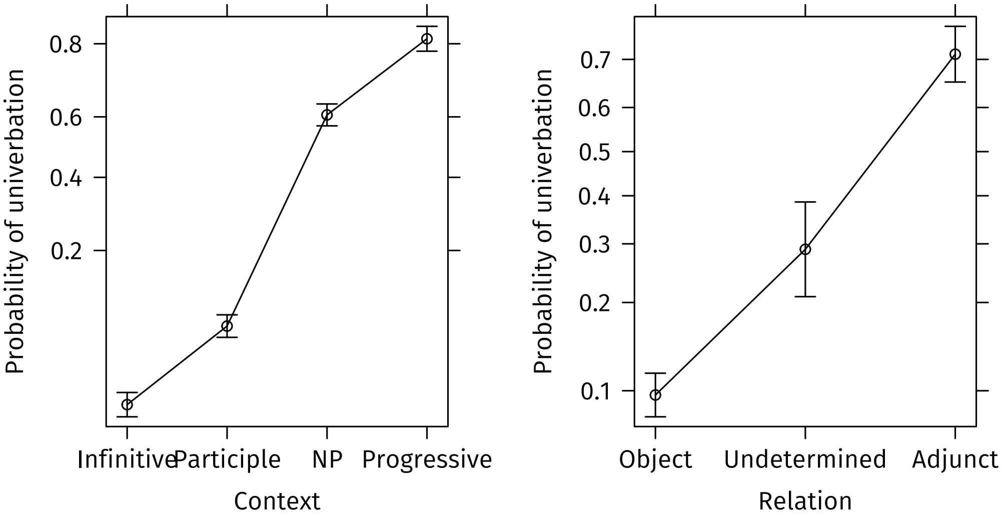

Figure 2: Effect plots for the regressor encoding the morphosyntactic context of the N+V unit and the regressor encoding the syntactic relation within the N+V unit in the GLMM modelling the corpus data.

Given the narrow confidence intervals and the high marginal measure of determination $R^2_m = 0.577$, we consider the hypotheses regarding fixed effects as well corroborated by the data, especially the effects of the context and the internal relation. The differences differences between specific

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

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).[36] The conditional modes are centred around a second-level intercept of 4.430 with a standard deviation of 2.105. 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 V + N unit thus plays a major role in writers' tendency to univerbate V + N units, which conforms the results from Section 3.2. The results obtained from that method will be used to predict participants' behaviour in the controlled experiment reported in Section 4.

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 corroborate the findings from the corpus study 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. 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 contexts. For each of the four contexts, we chose one N + V unit with a high and one

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

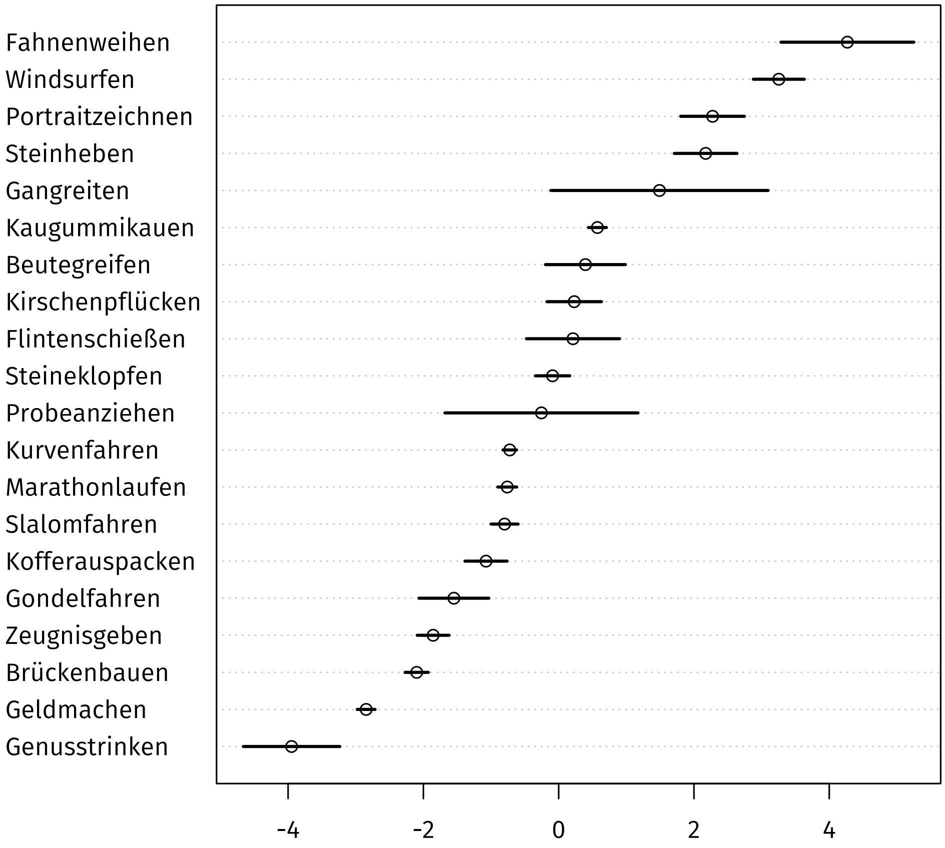

Figure 3: A random selection of conditional modes with 95% prediction intervals for the levels of the random effect in the GLMM modelling the corpus data.

with a low attraction strength according to the analysis of the usage data.[37] The sentences were constructed in a ways 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 full sentences are given in Appendix C.

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 made sure that there were at least three sentences in between pairs of items. There were nine distractors in the form of yes-no questions related to random sentences previously heard by the participants. An overview of the item design is shown in Table 5, where each line represents the features of one of the eight items.

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

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

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 stimuli was different between the two groups.

[37] Given the overall constraints on the choice of the items, “low” and “high” had to be interpreted as quite relative terms. However, we made sure that all low attractions scores are lower than zero and all high attraction scores are greater than zero. Also, for each context, the low and the high attraction score differ by at least 0.05.

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

4.2 Results

To analyse the results further, we report the parameter estimates of a GLMM with additional control for individual participants in the form of a random effect. Instead of using a grouping variable for the N + V units, we included their association strengths directly in the model. The model specification in R notation is given in (15). Appendix B provides the specification mathematical notation. The coefficient estimates for the GLMM are reported in Table 6.

(15) $\texttt{Univerbation} \sim (1|\texttt{Participant}) + \texttt{Attraction} + \texttt{Context}$

	Estimate	CI low	CI high
(Intercept)	-4.026	-5.533	-2.851
Attraction	48.740	34.657	73.466
Context = Participle	1.126	-0.375	2.574
Context = Progressive	6.224	4.691	8.184
Context = Clitic	8.166	5.978	11.512

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

There is some variation between writers as captured in the standard deviation of the conditional modes (1.722), but the small difference between the marginal R^2 (0.803) and the conditional R^2 (0.896) suggests that speaker variation does not explain much of the variance in the the data. The coefficients indicate that the attraction strength derived from the corpus is positively correlated with the participants' tendency to univerbate (48.740). 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 ($[-0.375..2.574]$). On the other hand, progressives (6.224) and NPs with cliticised articles (8.166) clearly have a much more positive effect on the probability of univerbation. Thus, in the GLMM analysis NPs appear to have a stronger tendency to favour univerbation than progressives. This is in line with our theoretical predictions.

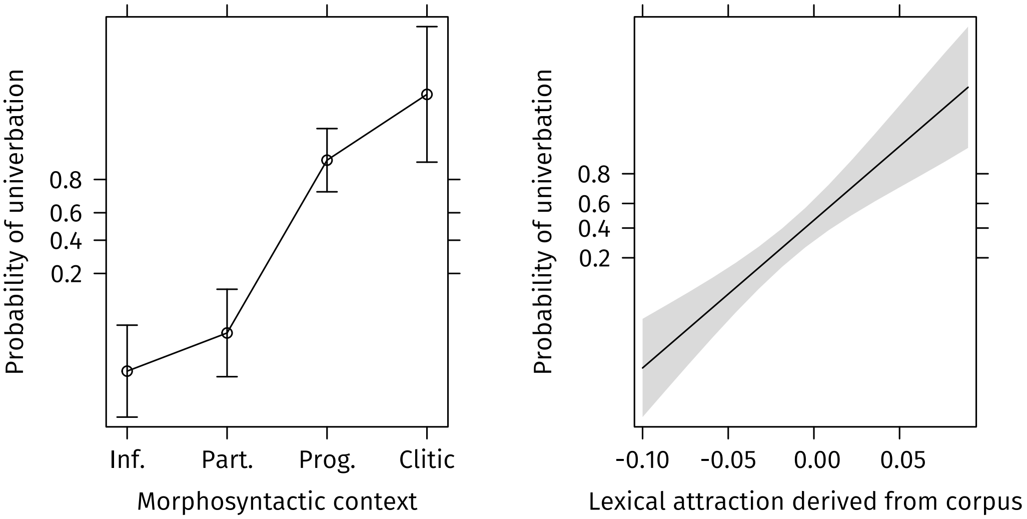

Figure 4: Effect plots for the regressor encoding the morphosyntactic context and the attraction strength as calculated from the corpus in the GLMM modelling the experimental data.

The effect plots in Figure 4 show the same picture as the coefficient table. The prototypically verbal contexts are associated with low probabilities of univerbation, the two prototypically nominal ones with high probabilities of univerbation. While progressives and NPs with clitics show the order predicted by theory, there is only very weak to no support for assuming a substantial difference judging by the large and mostly overlapping confidence intervals. The attraction scores are neatly correlated with the probability of univerbation.

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 hand in Section 5.

5 Explaining noun-verb univerbation

A Corpus study: full specification of the model

In Section 3.3, the specification of the model was given in R notation as (14), repeated here as (16).

(16) $$\texttt{Univerbation} \sim (\texttt{1}|\texttt{NVUnit}) + \texttt{Context} + \texttt{Relation} + \texttt{Link}$$

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 (16) has to offer. Mathematically and thus more transparently, the model is given in (17). 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.

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

The proportion of compound spellings $Prop_{Comp_i}$ is the logit-transformed sum of the overall intercept α_0, the random intercept for the j-th N + V unit $\alpha_{NV_j[i]}$ (whichever is found in observation i) and the dot product of the vector of dummy-coded binary value for the morphosyntactic context $\vec{x}_{Context_i}$ and the vector of their corresponding regressors $\vec{\beta}_{Context}$. Since it is a multilevel model, α_{NV} has its own linear model, which is given in (18).

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

It is also assumed that 19 holds, i. e., that the random intercepts for individual N + V units are normally distributed.

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

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

B Experiment: full specification of the model

In this appendix, we provide the mathematical notation of the model specified in Section 4.2 as (15) and repeated here as (20).

(20) $\texttt{Univerbation} \sim (\texttt{1}|\texttt{Participant}) + \texttt{Attraction} + \texttt{Context}$

The model is specified in the same notation as in Appendix A in (21). The regressor $x_{Attract_i}$ is numeric (the attraction score), whereas $\vec{x}_{Context_i}$ is a dummy-coded vector of binary variables.

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

It is also assumed that 19 holds, i. e., that the random intercepts for individual participants are normally distributed.

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

C Sentences used in the experiment

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

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

(24) 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.

(25) 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.

(26) 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.

(27) 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.

(28) 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.

(29) 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.

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

Acknowledgments

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

References

Arppe, Antti & Juhani Järvikivi. (2007). Every method counts: combining corpus-based and experimental evidence in the study of synonymy. *Corpus Linguistics and Linguistic Theory* 3(2). 131–159. http://dx.doi.org/10.1515/cllt.2007.009.

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.

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.

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.

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.

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.

Eisenberg, Peter. (2020). *Grundriss der deutschen Grammatik: Das Wort*. 5th edn. Stuttgart: Metzler.

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

Fleischer, Wolfgang & Irmhild Barz. (2012). *Wortbildung der deutschen Gegenwartssprache*. Marianne Schröder (ed.). 4th edn. Berlin, Boston: De Gruyter.

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 langauges of Europe*, 594–610. De Gruyter Mouton.

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.

Fuhrhop, Nanna. (2007). *Zwischen Wort und Syntagma. Zur grammatischen Fundierung der Getrennt- und Zusammenschreibung*. Tübingen: Niemeyer.

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.

Gaeta, Livio & Barbara Schlücker (eds.). (2012). *Das Deutsche als kompositionsfreudige Sprache: strukturelle Eigenschaften und systembezogene Aspekte*. Berlin: De Gruyter.

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.

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.), *Corpus-based approaches to syntax and lexis*, 159–191. Mouton De Gruyter.

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.

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.

Hentschel, Elke & Harald Weydt. (2003). *Handbuch der deutschen Grammatik*. 3rd edn. Berlin, Boston: De Gruyter.

Hoberg, Ursula. (1981). *Die Wortstellung in der geschriebenen deutschen Gegenwartssprache*. Vol. 10 (Heutiges Deutsch. Linguistische Grundlagen. Forschungen des Instituts für deutsche Sprache). München: Hueber.

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.

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.

Lehmann, Christian. (2021). Univerbation. *Folia Linguistica Historica* 42. MISSING.

Mithun, Marianne. (1984). The evolution of noun incorporation. *Language* 60(4). 847–894.

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.

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.

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.

Pauly, Dennis Nikolas & Guido Nottbusch. (2020). The influence of the German capitalization rules on reading. *Frontiers in Communication* 5(15). 1–15.

Primus, Beatrice. (2010). Strukturelle Grundlagen des deutschen Schriftsystems. In Ursula Bredel, Astrid Müller & Gabriele Hinney (eds.), *Schriftsystem und schrifterwerb*, 9–45. Berlin, New York: De Gruyter.

Rosch, Eleanor. (1973). Natural categories. *Cognitive Psychology*. 328–350.

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.), *The practical handbook of corpus linguistics*, 535–561. Berlin, Heidelberg: Springer.

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

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

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/10.1515/9783110278439.1.

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.

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.

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.