Revision accepted for publication by Written Language and Literacy on 19 September 2016.

Punctuation and Syntactic Structure in *Obwohl* and *Weil* Clauses in Nonstandard Written German

Roland Schäfer (roland.schaefer@fu-berlin.de)

Ulrike Sayatz

Deutsche und niederländische Philologie

Freie Universität Berlin

Habelschwerdter Allee 45

14195 Berlin

Abstract: In this paper, we analyze written sentences containing the German particles *obwohl* ("although") and *weil* ("because"). In standard written German, these particles embed clauses in verb-last constituent order, which is characteristic of subordinated clauses. In spoken and—as we show—nonstandard written German, they embed clauses in verb-second constituent order, which is characteristic of independent sentences. Our usage-based approach to the syntax–graphemics interface includes a large-scale corpus analysis of the patterns of punctuation in the nonstandard variants that provides clues to the syntactic structure and degree of sentential independence of the nonstandard variants. Our corpus study confirms and refines hypotheses from existing theoretical approaches by clearly showing that writers mark *obwohl* clauses with verb-second order systematically as independent sentences, whereas *weil* clauses with verb-second order are much less strongly marked as independent. This work suggests that similar corpus

studies could provide deeper insight into the interplay between syntax and graphemics.

Keywords: punctuation, syntax–graphemics interface, independent sentences, connectors, discourse markers, nonstandard writing, Prototype Theory, usage-based linguistics, German

Punctuation and Syntactic Structure in *Obwohl* and *Weil* Clauses in Nonstandard Written German

Abstract: In this paper, we analyze written sentences containing the German particles *obwohl* ("although") and *weil* ("because"). In standard written German, these particles embed clauses in verb-last constituent order, which is characteristic of subordinated clauses. In spoken and—as we show—nonstandard written German, they embed clauses in verb-second constituent order, which is characteristic of independent sentences. Our usage-based approach to the syntax–graphemics interface includes a large-scale corpus analysis of the patterns of punctuation in the nonstandard variants that provides clues to the syntactic structure and degree of sentential independence of the nonstandard variants. Our corpus study confirms and refines hypotheses from existing theoretical approaches by clearly showing that writers mark *obwohl* clauses with verb-second order systematically as independent sentences, whereas *weil* clauses with verb-second order are much less strongly marked as independent. This work suggests that similar corpus studies could provide deeper insight into the interplay between syntax and graphemics.

1. Introduction

In the existing literature on German nonstandard verb-second clauses headed by subordinating and/or coordinating particles (e.g., Antomo and Steinbach 2010, 2013; Reis

2013), graphemic evidence (especially punctuation) plays next to no role. Most prominently discussed are the particles *obwohl* ("although") and *weil* ("because"). In standard written language, these typically embed a clause in verb-last constituent order (VL), which is the typical order in embedded sentences. Additionally, these particles have nonstandard variants that embed clauses in verb-second constituent order (V2).[1] Although V2 is otherwise strongly (but not exclusively) characteristic of independent sentences, a major question is to what degree the V2 variants syntactically form completely independent sentences. In this paper, we present graphemic evidence—most prominently spontaneous use of the comma in nonstandard writing—showing that the V2 variants have a stronger tendency to form independent sentences than the VL sentences, and this tendency is much stronger for *obwohl* in V2 clauses than for *weil* in V2 clauses.

We now begin by demonstrating that the V2 variants are not, as is often assumed, exclusive to spoken language (cf. references in Section 2.1.2), and certain nonstandard patterns of punctuation can be found in the written V2 variants. We introduce examples of the relevant structures taken from the DECOW12Q web corpus (see Section 3 for details about the corpus and the Appendix for a list of the original URLs from which the examples were taken) for *obwohl* (1) and *weil* (2) in VL (a) and V2 (b).[2]

---

1 In syntactically subordinated clauses, *obwohl* and *weil* are usually called *subjunctors*. In the non-subordinated cases, they are often classified as *connectors* or *discourse markers*. We mostly call them *particles*, taking an agnostic stance with regard to a precise classification.

2 When inspecting the examples, notice that the primary use of the comma is very different in German compared to English in that there is a well-defined mapping from syntactic constructions to commas.

(1) a. Also ich bleib bei meinen George , obwohl
well I stay with my George , although
Arashi auch ziemlich lustig ist !
Arashi also rather funny is !
I still prefer George although Arashi is also rather funny!

b. Ich habs mir gegeben , obwohl am
I have.it me given , although on.the
Sonntag kamen manchmal wiederholungen vom
Sunday came sometimes repeats of.the
Samstag …
Saturday …
I watched [*all of*] *it, even though on Sunday they also showed some repeats from Saturday.*

Relative clauses (both restrictive and nonrestrictive), adverbial clauses, and complement clauses are obligatorily separated from their matrix clause by commas. The same is true for certain control infinitives that have a clausal status. On the other hand, integrated sentence-initial sentence adverbial phrases are usually not separated by commas (in contrast to adverbials like *on the other hand* in English). However, in nonstandard language, extraposed non-integrated adverbials and particles are separated from the rest of the sentence by a comma (cf. especially Section 2.2.3).

(2) a. Verschenken geht nur bedingt , weil das
make.present goes just limited , because that
ja nicht jedem gefällt ;-)
yes not everybody pleases ;-)
It doesn't make a good present either because many people don't like it.

b. Ich dachte nur ich komm an den DSLAM
I thought only I come to the DSLAM
da beim alten Kino , weil
there at.the old cinema , because
sonst steht hier näher keiner .
else stands here closer none .
I just thought I could get reception from the DSLAM by the old cinema. After all, there is no other access point in the vicinity.

The examples in (1) and (2) show that both particles occur with both VL and V2.[3] In (1) and (2), the *obwohl* and *weil* clauses follow their matrix clauses. However, the VL variants can also precede their matrix clause, as in (3a). Preposed V2 clauses, on the other hand, are ruled out. In (3a), for example, using the V2 alternative for the *weil* clause (*weil der Hang ist so steil*) would lead to a sentence with extremely low acceptability.

3 In the remainder of the paper, we use *weil-VL* for *weil* clauses with verb-last constituent order, and in a similar fashion *weil-V2*, *obwohl-VL*, and *obwohl-V2*.

We did not find a single sentence like this in our corpus sample (cf. Section 3) and consequently do not discuss this type of structure further. Thus, if a sentence begins with an *obwohl* or *weil* V2 clause, it always forms a fully independent sentence, as in (3b).

(3) a. Weil der Hang so steil ist , sind überall
because the slope so steep is , are everywhere
lauter Trockenmauern im Hang .
many dry.stone.wall in.the slope .
Because the slope is so steep, they have built many dry stone walls into it.

b. Weil das kann ich so wirklich nicht sehen .
because that can I so really not see .
Because[*, honestly,*] *I don't see that.*

It is an accepted fact that the V2 variants have a wider range of interpretations than the VL variants. In (2b), for example, a propositional causal interpretation for *weil* is excluded. The speaker's thinking that he might get connected to the access point is surely not caused by the fact that there are no other access points in the area. There are alternative epistemic and speech act–related readings of *weil* and *obwohl* in addition to their older propositional readings.[4] We do not go deeply into the complicated details of the

4 As for *weil* (as a discourse marker), Günthner (1996) calls it a marker of an *epistemic reason* or a *change of the discourse topic* (our translations). Günthner (2000) attributes to the discourse marker *obwohl* the function of a *disagreement marker* (our translation).

interpretation of the V2 variants here because the semantics and pragmatics do not play a major role in our corpus study (see Section 2.1 for the corresponding argument).

Finally, in (4), we show examples with punctuation marks (henceforth PMs) after the particle—in this case, the ellipsis in (4a) and the colon in (4b). As we demonstrate in Section 3, this use of PMs is typical of the V2 variants, and *obwohl*-V2 favors such PMs much more strongly than *weil*-V2.

(4) a. Oder ich könnte das Altmetall verwerten ,
or I could the scrap.metal use ,
obwohl … viel Metall ist da nicht
although … much metal is there not
dran .
at .
Or I could recycle it as scrap metal. But then again, it doesn't contain much metal.

b. wohin , das sag ich nicht , weil :
where , that say I not , because :
das weiß ich noch nicht .
that know I yet not .
I'm not going to say where, [*simply*] *because I don't know yet.*

In the larger picture, the goal of our paper is to show how very large corpora containing billions of tokens of nonstandard writing provide strong empirical support for usage-based inferences about the syntax–graphemics interface. The question of the degree of

sentential independence of *obwohl*-V2 and *weil*-V2 is used as an example, mostly because it has been discussed extensively in the theoretical literature—but so far without looking at usage data. The remainder of this paper is structured as follows. In Section 2, we briefly summarize some of the theoretical work on embedded V2 clauses in German. We also introduce our method of corpus-driven usage-based graphemics, and we develop our main hypotheses for the corpus study. In Section 3, we present the corpus study, examining the systematic syntactico-graphemic differences between *obwohl* and *weil* on the one hand and VL and V2 on the other hand. Finally, we summarize and interpret the findings in Section 4.

2. Theoretical Background

In this section, we introduce the theoretical background for our corpus study. We first review nongraphemic approaches to sentential independence and previous analyses of *obwohl* and *weil* clauses in Section 2.1. We then turn to graphemics in Section 2.2, focusing on the German punctuation system, our corpus-driven empirical approach, and the notion of *independent graphemic sentences*. In Section 2.3, we summarize the hypotheses for our corpus study.

2.1 Research on German Clause Types and Sentential Independence

2.1.1 Structural Integration in Previous Approaches

As pointed out in Section 1, we are interested in measuring the degree of sentential independence of *obwohl*-V2 and *weil*-V2 clauses. In the German theoretical literature on *obwohl*-V2, *weil*-V2, and similar phenomena, this is often discussed under the labels of

structural integration versus *structural nonintegration*. In this section, we argue that the operationalization of the distinction between integrated and nonintegrated clauses as used in the theoretical literature (and the corresponding tests) does not apply in large-scale corpus studies such as ours. In Section 2.2.3, we therefore propose a different approach to sentential independence that is rooted in cognitive linguistics (specifically, Prototype Theory) and rests upon a graphemic operationalization.

Prominently, Reich and Reis (2012) define and illustrate *subordination* and *coordination* based on their definition of *structural integration* (*strukturelle Integration*, p. 537). One problem with approaches rooted in syntactic and semantic theory, as the authors state themselves (Reich & Reis 2012: 542), is that reasonably definitive categorizations can only be given within specific theories. For example, Reich and Reis (2012: 551) state as the major property of nonintegrated clauses (such as coordinated clauses) that they are neither selected by the matrix as a complement clause nor related to it through a modification relation (such as adverbial or attributive clauses). Clearly, such definitions depend on assumptions about a specific version of phrase structure–based grammar and some theory of compositional semantics and the syntax–semantics interface.[5] Consequently, some of the tests that come with these theoretical notions are purely syntactic (e.g., across-the-board movement, p. 549), but many are also related to the syntax–

5 Even in their introduction, Reich and Reis (2012: 536–543) refer to notions from such diverse areas as the topological model of German sentence structure (p. 536), from Government and Binding Theory (c-command, p. 537), and semantic type theory (pp. 539–540).

semantics interface, such as pronoun binding (p. 537) or compatibility with modal particles (p. 551).

Per se, the dependence on specific theoretical frameworks would not stand in the way of an operationalization of the crucial notions. Also, we do not wish to dispute the value of the aforementioned tests for linguistic theorizing. However, their nature makes them fundamentally incompatible with corpus linguistic methods because they usually involve some modification of the attested sentence and a potentially difficult and problematic subjective judgment of the result of the modification. Standardly, one has to decide whether the modified sentence is still grammatical or whether it is semantically equivalent to the original sentence. Therefore, we clearly cannot rely on the established operationalizations of integration and nonintegration from theoretical linguistics in a corpus-driven study.

On the conceptual side, Reich and Reis (2012: 559–560) argue that syntactic integration is not a gradient notion but rather a categorical distinction connected to clearly distinguishable clause types. As is more appropriate for corpus-driven work, we approach the problem differently based on the rich support (gathered in cognitively oriented linguistics over the past decades) for the fact that grammar is inherently graded and probabilistic (Manning 2003; Hay & Baayen 2005; Bresnan 2007; Kapatsinski 2014, just to name a few publications). Apart from our general cognitive focus, usage data from corpora simply leave us no choice but to adopt a probabilistic interpretation. The data shown in Section 3 form clearly interpretable patterns. However, these patterns are not distributed categorically, and they always involve a random component. Therefore, although we will refer back to the notion of syntactic integration in the form of our discus-

sion of independent sentences in Section 2.2.2, we will do so from a cognitive linguistic perspective.

2.1.2 Views on *Obwohl*-V2 and *Weil*-V2

In our study, we focus on graphemic evidence for sentential independence. The degree of sentential independence of *obwohl*-V2 and *weil*-V2 clauses is, however, deeply related to their semantic, pragmatic, and prosodic properties. There is no reason to assume that graphemic evidence is of lesser value than, for example, prosodic evidence, which features prominently in the existing literature. To illustrate the relevance of our study in the context of the research on *obwohl*-V2 and *weil*-V2, we now briefly sketch some of this research.

In general, the V2 variants are treated as *noncanonical clausal constructions* (*nicht-kanonische Satzkonstruktionen*, Holler 2009: 135) restricted to spoken German (cf. Gaumann 1983; Günthner 1993; Wegener 2000; Pasch 1997; Uhmann 1998; Antomo & Steinbach 2010, 2013; Reis 2013). There are some empirical accounts of the phenomenon that exclusively use sparse data from corpora of spoken language (cf. Gohl & Günthner 1999; Freywald 2010: 60). However, we are not aware of existing larger studies of *obwohl*-V2 or *weil*-V2 in spoken German, except for that by Volodina (2011), which is discussed later in the paper. *Obwohl* is sometimes assumed to have a well-established secondary function as a discourse marker already (e.g., Günthner 2000),

whereas for *weil*, the development is said to be still ongoing (e.g., Gohl & Günthner 1999; Günthner 2003).[6]

With regard to the syntactic status (in the sense of the discussion in Section 2.1.1), Reich and Reis (2012: 557–558) analyze the VL variants as subordinate and thus integrated, and they describe V2 variants as nonintegrated. A related question is how strongly VL and V2 order are mapped onto the different readings, and it appears that constituent order alone is not a reliable indicator. Holler (2009: 136) and Blühdorn (2008: 217) show convincing examples of *weil*-V2 with causal interpretations. Also, noncausal readings for *weil*-VL have been demonstrated (Wegener 2000: 69; Volodina 2011: 82–83). Reis (2013: 228) concludes that syntactically nonintegrated *weil* clauses show V2 order by default, but not exclusively.[7]

With regard to prosody, Antomo and Steinbach (2010: 9) argue that nonintegrated clauses form their own intonational unit. Similarly, and with reference to empirical findings from Volodina (2011), Reis (2013: 229) attests that it is characteristic of nonintegrated clauses to be intonationally separate.[8] Blühdorn (2008: 5) finds that syntactic nonintegration is often (but not always) accompanied by a separate prosodic phrasing of the connected sentences. Gohl and Günthner (1999: 46–47) make a slightly different

6 For further discussion of diachronic data, see Freywald (2008, 2010).

7 For a discussion of nonintegrated VL clauses headed by *obwohl*, see Antomo and Steinbach (2013: 446).

8 In this context, nonintegrated and thus possibly noncausal *weil* VL clauses can behave similarly. Pasch (1983: 332–333) assumes that in order for *weil* VL to be interpreted as noncausal, an intonation boundary between the *weil* clause and the matrix is obligatory (cf. Volodina 2011: 83).

claim, stating that even the particle *weil* itself can form an intonational unit of its own. Auer and Günthner (2005: 341–342, 336) argue that the *obwohl* and *weil* in V2 are discourse markers that are syntactically and semantically *distant* from the rest of the sentence compared with their use as subjunctors in VLs. This distance can (but does not have to) be marked prosodically by pauses. Breindl (2009: 277) assumes that prosodic nonintegration is marked by pauses before and after *obwohl* in V2 but that *weil* in V2 is not marked by a following pause. Because native speakers will readily confirm that the pause after *weil* is quite admissible, this claim might be too strong, however. In an empirical study, Volodina (2011:159) concludes, on the basis of prosodic cues, that *weil*-V2 has a low degree of syntactic integration. As she diagnoses, however, there is no perfect mapping from syntactic integration to prosodic integration (Volodina 2011: 223).

To summarize, it seems to be an accepted fact that V2 is neither necessary nor sufficient in order to make the nonpropositional readings available. Specifically, the discussions of nonintegrated VL clauses with *weil* in Reis (2013) and of those with *obwohl* in Antomo and Steinbach (2013) make it clear that prosodic factors and certain speech act–related particles support the functional reinterpretation of the subjunctor and the subsequent shift in reading. Prosodic nonintegration, especially, is often treated as a de facto requirement (Reis 2013: 243; Antomo & Steinbach 2013: 446–447).

Finally, we turn to some scattered remarks on punctuation in nonintegrated clauses. Fahrländer (2013: 9) interprets colons and dashes—but not ellipses, as in the previous example (4a)—after *obwohl* as indicators of intonational pauses. A similar interpretation is advocated by Pasch et al. (2003: 406) for commas, colons, and dashes after

weil in V2 clauses, as in (3b). Antomo and Steinbach (2013: 427–428) illustrate formal and functional differences between types of *obwohl* clauses by marking them with different PMs, implicitly interpreting them as indicators of prosodic integration or nonintegration. This becomes most evident in their discussion of their example (12), repeated here as example (5).

(5) Ich nehme das Auto . Obwohl : Es gibt
I take the car . Although : It gives
an der Uni keine Parkplätze .
at the university no parking.spaces .
I'll take the car. But then again, there are no parking spaces near the university.

Regarding (5), Antomo and Steinbach (2013: 435) give the following comment, alluding to the possibility of *seeing* an intonational unit:

> As can be seen in example 12, the two clauses before and after *obwohl* form two separate intonational units, and there is an intonation pause after *obwohl* [...].[9]

In Antomo and Steinbach's (2013) study about interpretations of written variants of *obwohl* clauses, participants were exposed to different patterns of punctuation. There is no

9 Wie in Beispiel 12 zu sehen ist, bilden die beiden Sätze vor und nach obwohl jeweils eine separate Intonationseinheit und auf obwohl folgt eine intonatorische Pause [...].

discussion of those patterns from a graphemic perspective. We repeat Antomo and Steinbach's (2013: 427–428) examples (1) and (2) here as (6) and (7). The glosses are ours.

(6) Ich komme mit ins Kino , obwohl ich noch
I come with to.the cinema , although I still
lernen muss .
study must .
I will come to the cinema although I still have to study.

(7) a. Ich komme mit ins Kino . Obwohl :
I come with to.the cinema . although :
Ich muss noch lernen.
I must still study .
I will come to the cinema. But then again, I still have to study.

b. Ich komme mit ins Kino . Obwohl
I come with to.the cinema . although
ich noch lernen muss …
I still study must …
I will come to the cinema. Although/But then again, I still have to study.

The difference between the canonical subordinate clause in (6) and the nonintegrated clause in (7b) is exclusively graphemic. However, Antomo and Steinbach (2013: 437) only discuss constituent order as a factor distinguishing the different structures. We do

not deny a connection between prosody and punctuation, but we consider it problematic that—without taking the graphemic research into consideration—participants of studies using written stimuli are confronted with similar material, and graphemic aspects are not taken into account in the design of the study. With that in mind, we proceed to a discussion of punctuation in German in the next section.

2.2 Graphemics and Graphemic Markers of Sentential Independence

2.2.1 Usage-Based Graphemics

We see two major positions in the theoretical analysis of punctuation in German, and both seem to agree inasmuch as they regard periods and commas—and to a lesser degree the colon, the dash, and other PMs—as being primarily related to syntactic structure rather than prosodic structure.[10] Established accounts (Mentrup 1983; Behrens 1989; Baudusch 1995; Gallmann 1996) as well as normative approaches (Augst et al. 1997) follow a *construction-based* interpretation, analyzing punctuation as a conventionalized mapping from syntactic constructions to PMs. Bredel (2008, 2011) calls this an *offline* analysis and contrasts it with an alternative *online* view. Based on an interpretation of some experimental research (Frazier & Rayner 1988; Mazuka & Lust 1990 and other references in Bredel 2008: Ch. 3), Bredel analyzes PMs in German as reading aids from a reader-centric perspective. It is not a matter of debate that PMs are important

10 For more details on the relationship between prosody and punctuation (also from a historic perspective), see Kirchhoff and Primus (2014: 196–198). On the diachronic shift from a prosodically motivated to a syntactically motivated punctuation system, see Nerius (2007: 236–241).

cues to linguistic structure for readers. Bredel's approach, however, does not readily explain the effects that we see in production data, simply because it lacks a writer's perspective.[11] In corpus data, we exclusively observe the results of the productive use of punctuation by writers. Any variation in nonstandard corpus data is, in our view, best modeled with reference to cognitive processes pertaining to the writer. Therefore, we adhere to the construction-based view and propose that writers' use of punctuation in syntactic constructions that are conventionally not used in writing (such as *obwohl*-V2 and *weil*-V2) is guided to a large extent by constructional similarity and a prototypicality mapping from syntax to graphemics. We label this approach *usage-based graphemics* and elaborate on it in the remainder of this section.

It is the common opinion that the two highly frequent PMs in German (period and comma) primarily correlate with syntactic structures. Therefore, we started out looking for cues to the degree of sentential independence of noncanonical *obwohl* and *weil* clauses by looking at the use of PMs before and after them (e.g., the examples in Section 1) and comparing them to similar uses of PMs in more conventional constructions. We base this method on widely supported concepts from cognitively oriented linguistics and usage-based grammar. In usage-based frameworks (see Bybee & Beckner 2009 for an overview), "language is seen as an inventory of dynamic symbolic conventions (constructions) whose organization is constantly updated by (and hence adapting to) language use" (Zeschel 2008: 1, referring to Langacker 2000). We assume that writers who spontaneously produce patterns of punctuation do so because they have repeatedly been

11 This was indirectly pointed out by Paschke (2010: 148).

exposed to such patterns co-occurring with morphological, syntactic, and even semantic and pragmatic patterns in written language. In usage-based theory, it is expected that the knowledge of language resulting from such repeated exposures consists of a complex network between such patterns and their parts. The network encodes both highly specific idiosyncratic pairings as well as very general productive schemas. The degree of productivity or idiosyncrasy is a function of the token and type frequencies encountered in usage data (Bybee & Beckner 2009: 832–842).

Some theories within the usage-based paradigm focus on the fact that speakers and writers make use of such networks and the similarity of new items to already known items when faced with the task of categorizing the new items. Technical details aside, such mechanisms are assumed and experimentally supported both in Prototype Theory (e.g., Rosch 1973; Rosch et al. 1976; Rosch 1978) and Exemplar Theory (e.g., Medin & Schaffer 1978; Hintzman 1986). Thus, we can expect that when a writer needs to find a way to put into writing a construction that he or she has never or rarely put into writing before, the simplest strategy is to apply the conventions (i.e., schemas learned through repeated exposure) from the most similar syntactic construction—that is, to map prototypical syntax onto prototypical punctuation (or graphemics in general).

After a first inspection of the data, we found that there are two aspects in particular that relate the use of PMs to sentential independence. The first one is, of course, the dedicated marking of sentential independence with appropriate PMs, which we discus in Section 2.2.2. The other one is a characteristic use of sentence-initial particles followed by PMs typical of independent sentences, which we discuss in Section 2.2.3.

2.2.2 Independent Sentences in Syntax and Graphemics

Any discussion of graphemic evidence of sentential independence requires at least a minimal discussion of what constitutes an independent sentence. Morphosyntactic, semantic, and pragmatic definitions of sentencehood are usually thought of as being riddled with problems (see, e.g., summary in Panther & Köpcke 2008: 85–88). However, it has been argued convincingly by Panther and Köpcke (2008) that the essence of sentencehood can be specified clearly if one accepts that the sentence is a nondiscrete and fuzzy prototypical category.[12] Prototypical properties of independent sentences in English according to Panther and Köpcke (2008: 94–96) include morphosyntactic and prosodic ones (such as a subject in the nominative and falling intonation) as well as semantic and pragmatic properties (such as assertive illocutionary potential). A similar list for German should obviously include V2 constituent order as a prototypical property of independent sentences, among others properties (cf. also Fabricius-Hansen 2011). We thus assume that the sentence as a syntactic unit does indeed have a cognitive reality in terms of Prototype Theory, and that speakers classify linguistic entities as syntactically independent sentences or not, be it more prototypical exemplars such as the one in (8a) or less prototypical ones as in (8b)–(8d).

12 It would in fact be surprising if the sentence as a category were more discrete than the numerous other categories that have been demonstrated to show prototypicality effects, such as parts of speech (Uehara 2003), inflectional classes (Schäfer 2016 aop), and syntactic constructions (Gries 2003; Divjak and Arppe 2013; Dobrić 2015).

(8) a. The cat sat on the mat.

b. Me a spy?

c. Nice weather.

d. Hello!

Furthermore, we consider it obvious that the sentence is also the domain in which a large class of constructions (usually called *syntactic constructions*) is instantiated. As such, it must have some cognitive reality, something that is not always taken for granted in the cognitively oriented literature. Grammatical frameworks—including cognitively oriented ones—often operate with a merely implicit notion of the sentence. Croft (2001), for example, refers to sentences throughout his book. However, there is no definition or even discussion of what a sentence is. In Croft (2004: 651), the author states that "[p]articular sentences instantiate constructions, usually multiple constructions." This is, in principle, compatible with our view. Langacker (2008: 481–483), being more specific, speculates that *clauses* map more or less well to what he calls *attentional frames* in spoken language but that

> [t]here is some truth to the view that segmentation into sentences is merely a convention of writing; a sentence is then definable (roughly) as a sequence bounded by spaces that begins with a capital letter and ends with a period. But segmentation is often arbitrary, [footnote omitted; RS/US] and many sequences written in this fashion are not traditionally considered sentences.

The fact that segmentation of sentences in writing is not fully deterministic does not necessarily mean that it is *arbitrary*. Denying the sentence as a linguistic category is

also formally inadequate because many syntactic phenomena are evidently not clause-bound and can only be explained if we assume an additional syntactic unit that is potentially larger than clauses. A very obvious such phenomenon is unbounded dependencies. Whereas simple long-distance dependencies (such as rightward extraposition) are by definition nonlocal but clause-bound, unbounded can span across many clause boundaries, but they are still sentence-bound dependencies (e.g., Pollard & Sag 1994: Ch. 4; Levine & Hukari 2006).

Based on a prototypical definition of sentences, we propose that there is a mapping from prototypical fully independent syntactic sentences to independent graphemic sentences, wherein *full syntactic independence* is for all of our purposes the same as *maximal nonintegration*. If a specific syntactic structure as produced by a writer has (in the mind of that writer) more of the prototypical properties of sentencehood, then he or she will be more likely to mark it as an independent sentence by graphemic means. Graphemically, full sentential independence is marked by specific sentence-ending PMs (combined with initial capitalization). We agree with a large number of researchers that sentence-ending PMs in German are the period, the question mark, and the exclamation mark (Dürscheid 2006: 153–154). Within a fully independent graphemic sentence, the comma is a weaker marker of independence. According to Primus (1993, 2010: 35–36), one of the two primary functions of German commas is to mark clause boundaries (subordination) and sentence boundaries (sentential coordination). Subordinated clauses intrinsically have a partially independent status. Sentence-coordinating commas, on the other hand, allow writers to explicitly mark two syntactically independent sentences as

less independent, probably for semantic and pragmatic reasons. Thus, we even expect a paradigmatic continuum of the prototypical use of PMs between two words in German:

1. No PM = full integration (subclausal constituent boundary)
2. Clausal comma = partial independence (clause boundary or boundary between independent sentences marked explicitly for reduced independence)
3. Period, exclamation, question mark = full independence (sentence boundary)

What stands out as obvious is that the in-between status of clausal commas is connected in a very specific way to the morphosyntactic, semantic, and pragmatic prototypical properties of sentences as discussed previously. Dependent clauses often have the morphosyntactic properties of prototypical sentences (such as having a subject). However, they virtually never have their pragmatic properties (such as illocutionary force). In German, V2 and VL order prototypically map to independent and subordinated sentences, respectively, in the syntax, but the embedded V2 phenomena introduced in the previous sections blur the boundaries and result in exemplars with difficult to assess prototypicality status. However, we can now make an obvious prediction, given our view of the syntax–graphemics interface: to the degree that *obwohl*-V2 and *weil*-V2 constitute nonintegrated independent sentences, they should occur more often as graphemically independent sentences, typically being preceded by sentence-ending PMs and not being preceded by a comma.

2.2.3 Punctuation After Sentence-Initial Particles

In Section 1—especially example (4)—we showed V2 examples with punctuation after *obwohl* and *weil*. As it turns out (see following discussion, esp. Table 1), the comma is by far the most frequent PM in this position. Primus (1993: 253, 2010: 35) subsumes these commas under the second primary function of commas, namely, *extraposition* in the broad sense. We propose that these commas, too, are a sign of sentential independence. As it turns out, there is a larger class of words used in such a way. The examples in (9) show a selection of German sentence-initial particles used with commas.[13]

(9) a.

Klar	,	der	Patient	kann	auch	einfach	2	Tabletten
sure	,	the	patient	can	also	simply	2	pills
nehmen […]								
take								
Sure, the patient could equally well just take two pills […]								

13 In their discussion of a superficially similar class of discourse particles, Pasch et al. (2003: 439–450) focus on those that *embed* an additional clause and form a hidden conditional clause that is itself embedded under a matrix clause (Pasch et al. 2003: 439), such as *angenommen* ("[if it is] assumed [that]") or *unterstellt* ("[if it is] presumed [that]"). Freywald (2016: 327), with reference to Pasch et al. (2003), puts these particles close to *obwohl* and *weil*. We propose that the conditional reading brought about by *angenommen*, *unterstellt*, and similar words and the propositional embedding relation set them apart from *ah*, *ach*, and *naja*, as well as *obwohl* and *weil*.

b. Andererseits , dieses Tuch ist umstritten
on.the.other.hand , this shroud is debated
.
.
On the other hand, [*the authenticity of*] *this shroud is under debate.*

c. Nun , dieser Anblick beweist , dass der
well , this sight proves , that the
männliche Penis eigentlich potthässlich ist .
male penis actually butt-ugly is .
Well, this sight proves that the male penis is actually butt-ugly.

d. Zugegeben , das sind die Highlights des
admittedly , that are the highlights of.the
Religionsunterrichts .
religious.education .
Admittedly, these are the highlights of religious education.

Table 1 provides an overview of the ten most frequent words occurring sentence-initially in DECOW12Q (Section 3.1) with the four most frequent PMs (colon, comma, dash, ellipsis) after them.[14] The PM that does not really fit in is the colon. It has a performative character and classifies the type of information encoded in the following sen-

14 Imo (2012) shows that a similar sentence-initial position can also be filled by larger syntactic objects. We restrict our discussion to syntactically simplex particles.

tence. The other three PMs have an overlap in terms of the sentence-initial words after which they occur. But the comma is the only one occurring with considerable frequency. When we compare the three PMs in this construction with their overall distribution in the corpus, it is not very unusual, however. The comma is slightly more frequent here than we would expect given its overall frequency, but the effect is small. A χ^2 test produces a Monte Carlo-simulated $p = 0.001$ ($n = 102{,}163{,}089$ with 1,000 replicates) but with a mere $V_{\text{Cramér}} = 0.037$. In other words, the comma is the prominent PM here, but no more prominent than it is everywhere else.

Table 1. The ten most frequent sentence-initial words separated from the rest of the sentence by a PM in DECOW12Q; the total count and the percentage among all sentences that begin with such a word are given; the total counts do not include hapaxes, and percentages are calculated relative to the given total counts.

Colon (total 1,244,898)			
Word	**Translation**	**%**	**Count**
PS	P.S.	6.84	85,147
Zitat	quote	5.51	68,600
Edit	edit	4.03	50,203
EDIT	edit	2.38	29,595
Wohnort	place of residence	2.23	27,719
Fazit	summary	2.12	26,364
P.S.	P.S.	1.91	23,725
Also	well	1.4	17,369
Beruf	profession	1.12	13,952
Aber	but, however	1.01	12,611

Comma (total 3,191,317)			
Word	**Translation**	**%**	**Count**
Ja	well, yes	7.5	239,380
Naja	well	6.21	198,089
Also	well, now	3.8	121,348
So	now	3.72	118,625
Nein	no	3.51	111,866
Tja	well	1.99	63,381
Sorry	sorry	1.83	58,403
Klar	obviously, yeah	1.64	52,447
Ok	okay	1.46	46,489
Gut	well	1.4	44,729

Dash (total 170,789)			
Word	**Translation**	**%**	**Count**
Und	and, furthermore	1.96	3,353
Also	well, now	1.72	2,940
Aber	but, however	1.59	2,711
Naja	well	1.49	2,548
Ja	well, yes	1.38	2,355
So	now	1.09	1,858
Nein	no	0.99	1,698
YouTube	YouTube	0.97	1,664
Tja	well	0.66	1,121
Klar	obviously, yeah	0.65	1,118

Ellipsis (total 210,593)			
Word	**Translation**	**%**	**Count**
Naja	well	5.46	11,488
Hm	hm	4.23	8,916
Hmm	hm	4.2	8,844
Also	well, now	2.91	6,119
Hmmm	hm	2.87	6,039
So	now	2.28	4,796
Aber	but, however	1.83	3,854
Ja	well, yes	1.82	3,823
Tja	well	1.76	3,700
Ähm	um	1.53	3,219

This type of comma (as a marker of extraposition in the sense of Primus 1993, 2010) corresponds to the strong syntactic and semantic disconnectedness of discourse markers from the rest of the sentence (Section 2.1.2; Auer & Günthner 2005; Breindl 2009) and

to the prosodic boundary that was reported in the previous literature.[15] Interestingly, Breindl (2009) argues that intonational pauses occur only after *obwohl* and not after *weil*, which we expect to manifest itself in the frequency of the use of PMs (cf. Section 3). Such sentence-initial particles followed by a comma usually require an independent root clause to attach to. Thus, if *obwohl* and *weil* should turn out to be systematically assimilated to this class graphemically (constructional similarity leading to graphemic assimilation as proposed in Section 2.2.2), it would be strong graphemic evidence for the nonintegration of these clauses. Although this use of PMs is related to intonation boundaries (cf. Section 2.1.2), it is syntactically motivated and cannot be reduced to a simple mapping from prosody to graphemics, especially because intonation boundaries are optional.

2.3 Summary and Hypotheses

In summary, we follow two main hypotheses. The first one is that *obwohl*-V2 and *weil*-V2 behave more like discourse markers than their VL counterparts, enforcing greater sentential independence (nonintegration). The second one is that *obwohl* has more prototypical properties of discourse markers in comparison with *weil*. Graphemically, we expect these hypotheses to be reflected in two ways. First, the V2 clauses should be marked more frequently as independent graphemic sentences with sentence-ending PMs. Second, the particles should be followed by sentence-internal PMs in V2

15 This corresponds well with the analysis by Primus (1993: 250–254), in which Primus also focuses on a strong connection between intonation and commas marking extraposition.

to the extent that they are construed as discourse markers. We expect both of these graphemic effects to be stronger in *obwohl*-V2 than in *weil*-V2.

3. Corpus Study

3.1 Choice of Corpus, Sampling, and Annotation

The type of corpus analysis performed in this work necessitates the use of large corpora because the phenomenon is rare in written language. Also, the corpora have to contain nonstandard variation. These requirements are met by large web corpora containing computer-mediated communication (forums, blogs, etc.). Such corpora were made popular by the WaCky initiative (Baroni et al. 2009). The similar but improved COW corpora were released in 2012 (Schäfer & Bildhauer 2012, 2013; Schäfer et al. 2013). They were evaluated as being of equal quality as reference corpora such as the British National Corpus (BNC) in collocation extraction tasks in Biemann et al. (2013). Only one year after their release, the corpora had been used in published research (e.g., Müller 2014; Schäfer & Sayatz 2014; Van Goethem & Hiligsmann 2014; Schäfer, in press). The DECOW12 corpus is 9.1 billion tokens large. Using a heuristic method, the corpus creators extracted a 1.8-billion-token subcorpus (DECOW12Q) that contains almost exclusively forum discussions and blogs. We used DECOW12Q.[16]

We exported two random samples containing 5,000 sentences, one for *obwohl* and one for *weil*. Removal of dubious (e.g., fragmentary) sentences reduced the samples to n = 4,739 (*obwohl*) and n = 4,784 (*weil*). The concordances were then annotated manually.

16 http://corporafromtheweb.org (information), https://webcorpora.org (download and query)

The annotation scheme shown in Table 2 captures the basic syntactico-graphemic distributional properties of the clauses in question. The annotation for *Hypo* was added in order to exclude cases in which the particle was followed by subordinated or parenthetic material (76 cases). In such clauses, any punctuation used after the particle is likely to be indicative of the syntax of the inserted material, as in example (10).[17]

(10) obwohl , ich weiß nicht warum , stehe momentan auf
Then.again , I know not why , stand recently on
Gold !
gold !
But then again, I don't know why, I've recently been fond of gold.

The annotation for *Mod* (i.e., whether there was an additional modifying particle to the left of the particle, as in *bloß weil* ["just because"]) was added because it was our impression that there were many such cases with *weil* but less with *obwohl*. Also, the modifier separates the connective particle from any PMs to its left, which might affect their use.

17 Pasch et al. (2003: 406–407) report higher acceptability for *weil* V2 clauses when there is a subordinated clause inserted directly after *weil*. Our data support this inasmuch as among the *weil* V2 sentences, we find 6.12% with *Hypo=1*, but among the *weil* VL sentences we only find 0.20% with *Hypo=1* (odd ratio 32.05, 95% confidence interval is [13.94, 80.04]). The effect is even stronger in the case of *obwohl*, with 14.18% versus 0.13% (odds ratio 122.2, 95% confidence interval is [50.87, 357.64]). However, the estimated odds ratios come with very large confidence intervals because of the low number of cases with *Hypo=1*.

Table 2 also includes count data for the levels of the annotated variables. Descriptively, the following results are remarkable. Although a comma (*Left=Comma*) is the most frequent left context for both particles, there is also quite a high number of occurrences after sentence-final punctuation (*Left=End*). The particles also occur without any preceding punctuation at all (*Left=Word*) as well as after other PMs, such as the ellipsis (*Left=Ellipsis*). After the particles, we mostly find words (*Right=Word*). However, commas after *obwohl* (*Right=Comma*) are also found (2.68%). Parenthetic material (*Hypo=1*) follows the subjunctors in under 1% of the cases. Modification of the subjunctors (*Mod=1*) indeed occurs more often with *weil* (13.42%) than with *obwohl* (5.13%).

As for constituent order, V2 (*Senttype=V2*) might appear to be rare at first sight, although roughly at the same level for *obwohl* (5.95%) and *weil* (7.17%). However, considering that this configuration is usually associated almost exclusively with *spoken* language (Gohl & Günthner 1999: 41; Antomo & Steinbach 2010: 1; Freywald 2010: 60–61; Schwitalla 2012: 142) and our data come from a corpus of *written* language, approximately 6% and 7% should rather be considered quite high percentages.

Table 2. Annotated factors, levels, raw count, and percentage in the sample. *n* (*obwohl*) = 4,739, *n* (*weil*) = 4,784.

Factor	Levels	Description	*obwohl*		*weil*	
			Count	**%**	**Count**	**%**
Left		**The particle follows:**				
	Comma	comma	2,329	49.15	2,868	59.95
	End	sentence-final punctuation (. ! ?)	750	15.83	1,132	23.66
	Word	word	972	20.51	568	11.87
	Paro	opening parenthesis (clause in parentheses)	287	6.06	35	0.73
	Ellipsis	ellipsis (three points)	170	3.59	71	1.48
	Dash	dash	97	2.05	43	0.90
	Emo	emoticon	97	2.05	42	0.88
	Colon	colon	37	0.78	25	0.52
Right		**The particle is followed by:**				
	Word	word	4,544	95.89	4737	99.02
	Comma	comma	127	2.68	1	0.02
	Ellipsis	ellipsis	42	0.89	7	0.15
	Colon	colon	9	0.19	26	0.54
	Dash	dash	17	0.36	13	0.27
Mod		**There is premodification of the particle.**				
	0	no	4,496	94.87	4,142	86.58
	1	yes	243	5.13	642	13.42
Hypo		**Subordinated or parenthetic material directly follows the particle.**				
	0	no	4,693	99.03	4,754	99.37
	1	yes	46	0.97	30	0.63
Senttype		**The embedded clause shows:**				
	VL	verb-last constituent order	4,457	94.05	4,441	92.83
	V2	verb-second constituent order	282	5.95	343	7.17

3.2 Modeling the Distributional Differences Between *Obwohl*, *Weil*, VL, and V2

In this section, we examine the syntactico-graphemic differences between VL and V2 in *obwohl* and *weil* clauses by means of estimating the coefficients of binomial generalized linear models (GLMs) with a logit link (i.e., logistic regression) using the annotated features described in Section 3.1.[18] To get the full picture, we divide the task into four different models. We look for distributional difference between *obwohl*-VL and *obwohl*-V2 in Section 3.2.1 and between *weil*-VL and *weil*-V2 in Section 3.2.2. In Section 3.2.3, we report the differences between *weil*-VL and *obwohl*-VL. We finally turn to the most important model in Section 3.2.4, which compares *obwohl*-V2 and *weil*-V2. Thus, we make sure that we discover all potential distributional differences between the four configurations (two particles and two clause types).[19]

3.2.1 *Obwohl* Clauses: VL Versus V2

The first model was specified so as to reveal distributional differences between *obwohl*-VL and *obwohl*-V2. The response variable was thus *Senttype* (see Table 2). Using the regressor *Left* straightforwardly was not possible because of the difficulty mentioned

18 Because introductory statistics textbooks for linguists that deal with GLMs (such as Baayen 2008; Johnson 2008; Gries 2013) have been available for many years, we do not provide an introduction to the method itself. We used the more complete Fahrmeir et al. (2013) and Zuur et al. (2009) as our main references.

19 Section 3.2.1 also contains elaborations on some basic procedures, which we do not repeat in Sections 3.2.2–3.2.4.

already in Section 1; see especially example (3) and the related discussion. If an *obwohl*-VL clause is positioned after a sentence-final PM, it can be a preposed (but otherwise perfectly integrated) adverbial clause with a matrix clause following it, or it can be—at least in nonstandard varieties—a truly independent sentence. An *obwohl*-V2 clause in the same position is never followed by a matrix clause and always forms an independent graphemic (and syntactic) sentence. In other words, *obwohl*-V2 clauses can never be preposed within a larger sentence, whereas after sentence-internal PMs (typically comma), both VL and V2 clauses can occur. We therefore annotated all cases with an additional variable *Independent*, which, in the case of VL, is 1 if there is no matrix following, and 0 otherwise. For all V2 clauses, *Independent* is 1 if a sentence-final PM precedes, and 0 otherwise.[20] In the GLM model structure, we also included an interaction term between *Independent* and *Mod* because we were not sure whether modification could influence the choice of constituent order (Section 3.1). The model was thus specified as

Senttype ~ Right + Independent * Mod

in R notation.[21] All cases annotated as *Hypo=1* were removed before estimating the models, which explains the slightly reduced sample size. For the model with estimated

20 For this, we extended the class of sentence-final PMs to include emoticons because after inspecting all uses of emoticons in the concordance, it appeared to us that they are used exclusively in positions where a period could also be inserted. Also, parentheses were included for similar reasons.

21 We use R notation throughout this section. For reasons of compactness, we also omit the equals sign for regressor levels in the tables and plots, such that *RightComma* should be read as *Right=Comma*,

coefficients, we used a step-down based on the AIC (Akaike Information Criterion) to remove uninformative regressors from the model specification. We do not consider a simple AIC-based step-down to be problematic in this case because more advanced methods of model selection, such as multimodel selection (used recently in the linguistics literature, for example, in Kuperman & Bresnan 2012 and Barth & Kapatsinski 2014, ahead of print), apply mainly to much more complex models (cf. Burnham & Anderson 2002: 04).

Table 3. Coefficient table of the binomial GLM for *obwohl* (logit link): V2 (positive coefficients) versus VL (negative coefficients) in *obwohl* clauses. Intercept: *Independent0*, *Mod0*, *RightWord.*

Regressor	***β***	**SE**	***z***	***p***	**Sign.**	**OR**
(Intercept)	−4.463	0.152	−29.40	<0.001	***	$1.153 \cdot 10^{-2}$
Independent1	2.058	0.225	9.13	<0.001	***	7.826
Mod1	−15.636	658.265	−0.02	0.98		$1.620 \cdot 10^{-7}$
RightColon	22.686	2545.826	0.01	0.99		$7.121 \cdot 10^{9}$
RightComma	8.191	1.018	8.05	<0.001	***	$3.608 \cdot 10^{3}$
RightDash	23.357	3618.083	0.01	0.99		$1.392 \cdot 10^{10}$
RightEllipsis	7.195	1.034	6.96	<0.001	***	$1.333 \cdot 10^{3}$

The estimated coefficients are reported in Table 3, where all interactions were removed by step-down. First of all, the model has an excellent Nagelkerke R^2 of 0.649, meaning that a great deal of the variance in the data is explained by the model. Because achieving significance levels is very easy with very large samples (n = 4,693), the high

and so forth. All model coefficients were estimated using R (R Core Team 2013) and the *glm* and *step* functions from the *stats* package. Additional packages used were *boot* (Canty & Ripley 2013) for the *cv.glm* function, *car* (Fox & Weisberg 2011) for the *vif* and *Boot* functions, and *fmsb* (Nakazawa 2014) for the *Nagelkerke* R^2 function.

R^2 is a much more meaningful measure than *p*-values.[22] The tenfold cross-validation error rate is 0.0171. This is a proportional improvement of 0.6712 (67.12%) over the baseline error rate of 0.052.[23] We checked for multicollinearity by calculating generalized variance inflation factors (GVIFs) for the regressors (Fox & Monette 1992): GVIF(*Independent*) = 1.005 (df = 1), GVIF(*Mod*) = 1.000 (df = 1), GVIF(*Right*) = 1.005 (df = 4). Diagnostic plots of the model residuals showed some degree of abnormal patterning and violations of homogeneity of variance, and we decided to run a bootstrap (Davison & Hinkley 1997; Fox & Weisberg 2011) with 10,000 replicates to get more robust estimates of the coefficients including bootstrap confidence intervals. The results are shown in Figure 1. The results are confirmed. The extended confidence intervals for *Right=Comma* and *Right=Ellipsis* are an indication that there are some abnormalities, but the signs of the coefficients are stable under bootstrapping. The contrasts that we found confirm our hypotheses from Section 2. Compared to *obwohl*-VL, *obwohl*-V2 occurs more often in graphemically independent sentences, and more often with commas and ellipses to the right of *obwohl*.

22 On the related topic of significance and weak effect strength, see the introductory comments in Baayen (2008: 114–116).

23 The baseline error rate is the error rate that one can achieve simply by always predicting the more frequent value of the dependent variable. In this case, it is the proportion of V2 clauses in the sample because V2 is slightly more frequent in the sample.

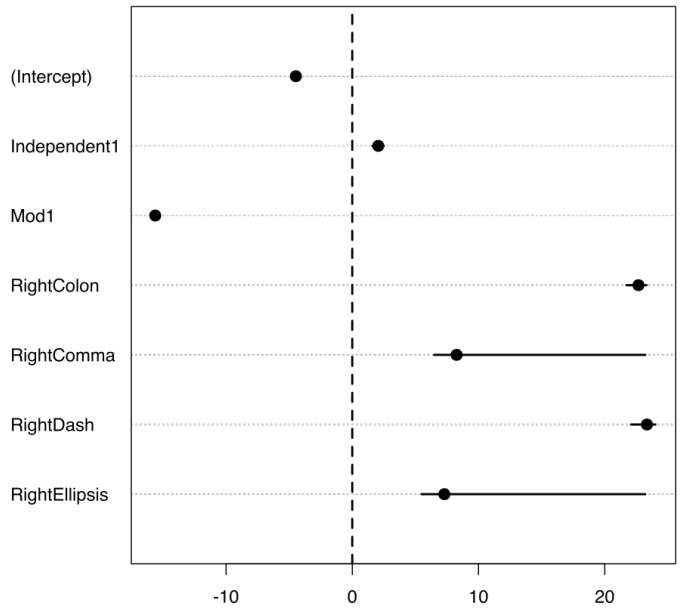

Figure 1. Bootstrapped estimates of the coefficients of the *obwohl* GLM (cf. Table 3) with 95% confidence intervals (10,000 replicates).

3.2.2 *Weil* Clauses: VL Versus V2

The procedure for the *weil* GLM (n = 4,754) was exactly the same as the one for the *obwohl* GLM, including the creation of the factor *Independent*, an identical model specification, and the removal of all cases with *Hypo=1*. Estimated model coefficients are reported in Table 4. The Nagelkerke R^2 of 0.181 is only marginally acceptable. There are no problems with collinearity: GVIF(*Independent*) = 1.008 (df = 1), GVIF(*Mod*) = 1.021 (df = 1), GVIF(*Right*) = 1.016 (df = 4), GVIF(*Independent: Mod*) = 1.000 (df = 1). Because of the low R^2, we decided not to interpret the model and skipped all further evaluations. Given the weakness of the model, we have only very weak evidence point-

ing toward informative distributional differences between *weil*-VL and *weil*-V2 in terms of full graphemic independence and PMs to the right of *weil*. This result is quite surprising because it is usually assumed that there are major structural differences between *weil*-VL and *weil*-V2 (see Section 2), and they should be expected to manifest themselves in the use of PMs.

Table 4. Coefficient table of the binomial GLM for *weil* (logit link): V2 (positive coefficients) versus VL (negative coefficients) in *weil* clauses. Intercept: *Independent0*, *Mod0*, *RightWord.*

Regressor	***β***	**SE**	***z***	***p***	**Sign.**	**OR**
(Intercept)	−2.911	0.074	−39.28	<0.001	***	$5.440 \cdot 10^{-2}$
Independent1	1.880	0.139	13.54	<0.001	***	6.551
Mod1	−1.957	0.478	−4.09	<0.001	***	$1.413 \cdot 10^{-1}$
RightColon	5.362	1.120	4.79	<0.001	***	$2.130 \cdot 10^{2}$
RightComma	18.598	3956.180	0.00	1.00		$1.194 \cdot 10^{8}$
RightDash	3.415	0.592	5.77	<0.001	***	$3.041 \cdot 10^{-1}$
RightEllipsis	19.951	1672.393	0.01	0.99		$4.619 \cdot 10^{8}$
Independent1: Mod1	−14.578	414.721	−0.04	0.97		$4.666 \cdot 10^{-7}$

However, the differences that we see in Table 4 still point in the expected direction. The signs of the coefficients make sense, and they point in the same general direction as those for *obwohl* (Section 3.2.1). In short, full graphemic independence and PMs to the right of *weil* occur proportionally more often in V2 clauses. The question of why premodifiers (see Section 3.1) co-occur proportionally more often with *weil* in VL clauses (*Mod=1*) than in V2 clauses cannot be answered here, although we assume it is related to semantic and pragmatic compatibility effects of *obwohl* and *weil* with such modifiers.

There is no interaction between PMs to the left of *weil* (here in the form of the *Independent* regressor) and premodifiers of *weil* as hypothesized in Section 3.1.

3.2.3 VL Clauses: *Obwohl* Versus *Weil*

With the VL GLM (n = 8,882), the goal is to model distributional differences between *obwohl* and *weil* in VL sentences. The model specification is as follows:

Particle ~ Right + Independent * Mod

Again, the regressor *Independent* had to be created because sentence-initial *weil*-VL can be followed by a matrix clause or not. Model estimates are shown in Table 5.

Table 5. Coefficient table of the binomial GLM for VL (logit link): *weil* (positive coefficients) vs. *obwohl* (negative coefficients). Intercept: *Independent0*, *Mod0*, *RightWord.*

Regressor	***β***	**SE**	***z***	***p***	**Sign.**	**OR**
(Intercept)	−0.072	0.023	−3.09	0.002	**	$9.302 \cdot 10^{-1}$
Independent1	−0.367	0.081	−4.53	<0.001	***	$6.931 \cdot 10^{-1}$
Mod1	1.015	0.084	12.04	<0.001	***	2.759
RightComma	−12.127	324.744	−0.04	0.970		$5.411 \cdot 10^{-6}$
RightDash	12.443	140.994	0.09	0.930		$2.534 \cdot 10^{5}$
RightEllipsis	−12.494	324.743	−0.04	0.969		$3.750 \cdot 10^{-6}$
Independent1: Mod1	0.501	0.237	2.11	0.035	*	1.651

In this case, the Nagelkerke R^2 of 0.035 is not even marginally acceptable. There are no problems with collinearity: GVIF(*Independent*) = 1.134 (df = 1), GVIF(*Mod*) = 1.147 (df = 1), GVIF(*Right*) = 1.000 (df = 3), GVIF(*Independent: Mod*) = 1.289 (df = 1). We do not interpret the model and did not run cross-validation or a bootstrap because of the very low R^2. In other words, we have found no evidence supporting the as-

sumption of informative differences between *weil*-VL and *obwohl*-VL in terms of full graphemic independence and PMs to the right of the two particles. This is an expected result because it has never been proposed that there are significant structural differences between *obwohl*-VL and *weil*-VL. Indirectly, this confirms that our data and our method of analysis produce meaningful results.

3.2.4 V2 Clauses: *Obwohl* Versus *Weil*

Finally, we turn to the most important model of the distributional differences between *obwohl* and *weil* in V2 clauses. Because V2 clauses after sentence-final PMs are always fully isolated (see Section 3.2.1), we could use *Left* as a regressor and did not have to use an aggregated regressor *Independent* as in the other models, leading to the following model specification:

Particle ~ Right + Left * Mod

The estimated coefficients of the GLM (n = 563) are shown in Table 6. *Mod* and the interaction term were removed by step-down. The Nagelkerke R^2 of 0.582 is excellent, and there is no serious collinearity: GVIF(*Left*) = 1.06 (df = 6), GVIF(*Right*) = 1.06 (df=4).[24] The tenfold cross-validation error rate is $\Delta = 0.1329$, which means a proportional reduction of error of 0.6896 (68.96%) over the baseline error rate of 0.4281.

24 Because there are only two regressors left in the model, the GVIF values are necessarily identical for both, which is a fundamental result following from Fox and Monette (1992).

Bootstrap estimates and confidence intervals (10,000 replicates) confirm the GLM estimates (cf. Figure 2).

Table 6. Coefficient table of the binomial GLM for V2 (logit link): *weil* (positive coefficients) versus *obwohl* (negative coefficients). Intercept: *LeftComma*, *RightWord.*

Regressor	***β***	**SE**	***z***	***p***	**Sign.**	**OR**
(Intercept)	1.879	0.235	8.00	<0.001	***	6.543
LeftEmo	−1.817	0.518	−3.51	<0.001	***	0.163
LeftEnd	−0.772	0.296	−2.61	0.009	**	0.462
LeftDash	−1.097	0.701	−1.57	0.117		0.334
LeftParo	−3.575	1.100	−3.25	0.001	**	0.028
LeftEllipsis	−2.719	0.579	−4.70	<0.001	***	0.066
LeftWord	−0.019	0.380	−0.05	0.961		0.982
RightColon	−1.355	0.435	−3.11	0.0018	**	0.258
RightComma	−5.645	1.017	−5.55	<0.001	***	0.004
RightDash	−0.663	0.535	−1.24	0.215		0.515
RightEllipsis	−3.103	0.506	−6.13	<0.001	***	0.045

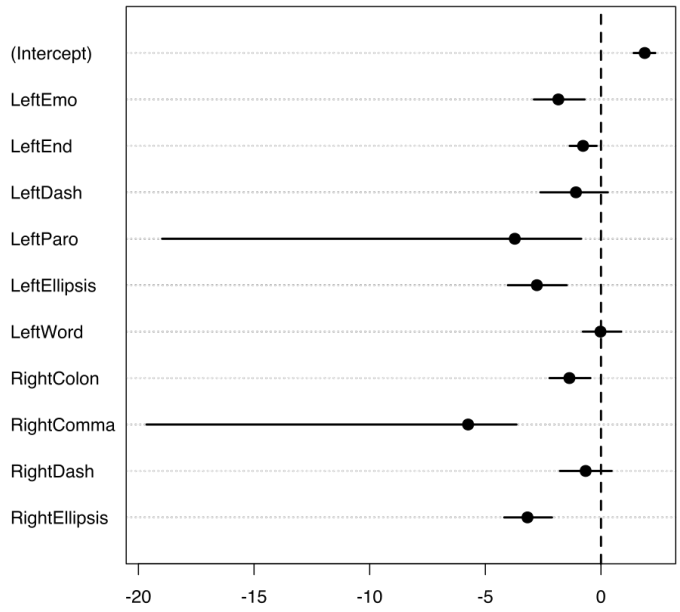

Figure 2. Bootstrapped estimates of the coefficients of the V2 GLM (cf. Table 6) with 95% confidence intervals (10,000 replicates).

The results are thus very clear, and they are visualized in a more straightforward manner in Figures 3 and 4, which show the distribution of the regressor levels in *obwohl* and *weil* clauses, including those with nonsignificant contrasts. Whereas *weil*-V2 proportionally prefers standard punctuation (*Left=Comma* and *Right=Word*, represented in the intercept in the GLM), *obwohl*-V2 co-occurs more often after sentence-final PMs, emoticons, and the ellipsis as well as within parentheses, and there are proportionally more cases with colons, commas, and dashes after *obwohl*. Thus, not only can we confirm that V2 clauses attract markers of full graphemic independence (Sections 3.2.1 and 3.2.2) and that PMs are used to mark the initial particles as discourse markers, but also

that *obwohl* has a much stronger tendency toward being marked in such a way. A linguistic interpretation of these results is provided in Section 4.

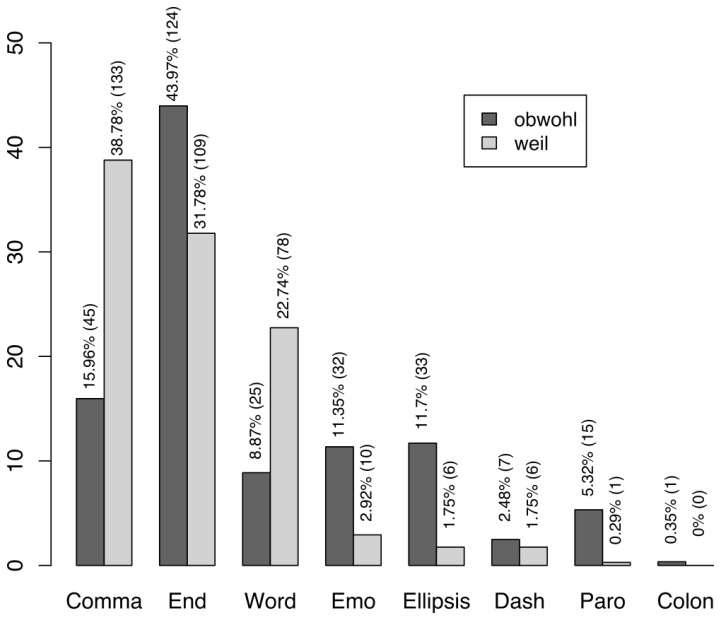

Figure 3. From the dataset for the V2 GLM (Table 6): Counts for the response variable *Subjunctor* (*obwohl* or *weil*) and the regressor *Left*.

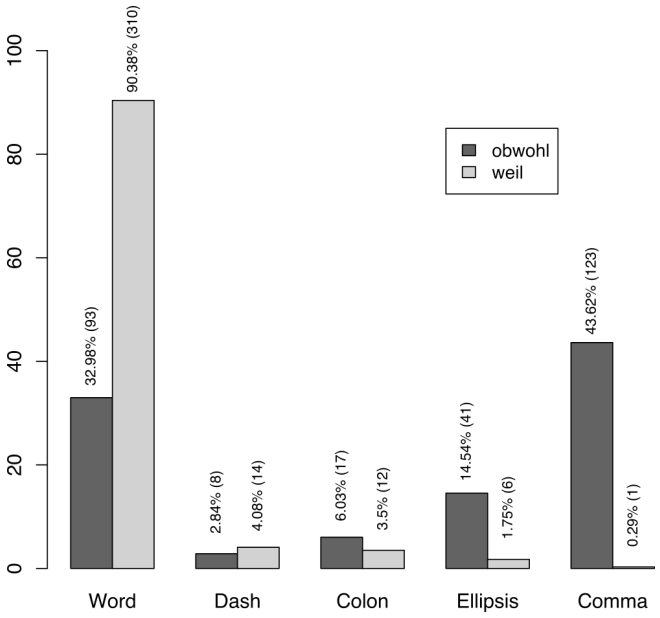

Figure 4. From the dataset for the V2 GLM (Table 6): Counts for the response variable *Subjunctor* (*obwohl* or *weil*) and the regressor *Right*.

4. Linguistic Interpretation and Outlook

In general, the hypotheses developed in Section 2 are supported by the data reported in Section 3. However, the differences between *obwohl*-V2 and *weil*-V2 were even bigger than we expected. Whereas *obwohl*-V2 often occurs with graphemic markers of independence, this is much less frequently the case for *weil*-V2. *Obwohl*-V2 tends to form independent graphemic sentences, as defined in Section 2.2.2, marked by sentence-final punctuation. It also behaves sentence-internally like a discourse marker, as illustrated in Section 2.2.3, through the co-occurrence with PMs to its right (predominantly the comma), which is a sign of nonintegration because discourse markers typically require an independent sentence to associate with. All this is much less strongly the case for *weil*, leading to only marginal contrasts between VL and V2 in *weil* clauses (Section 3.2.2).

Also, in direct comparison, *obwohl*-V2 shows much clearer preferences for markers of graphemic independence than *weil*-V2 (Section 3.2.4). At the same time, both *obwohl*-V2 and *weil*-V2 are equally well established in the DECOW12Q register (see Section 3.1, especially Table 2), and the differences cannot plausibly be attributed to *obwohl*-V2 being a less-well-established construction. In VL clauses, on the other hand, there are some differences between *obwohl* and *weil*, but they are extremely marginal and not worthy of interpretation.

Notice also that the continuum between full graphemic integration (no PM), partial integration (comma), and full nonintegration (sentence-final PM), which we proposed in Section 2.2.2, is also confirmed. Figure 3 shows that *weil*-V2 occurs more often without any PM to its left (*Left=Word*), namely, in 22.74% of all occurrences, compared with 8.87% in *obwohl*-V2.[25] This lends further support to the conclusion that *weil*-V2 is prototypically more integrated than *obwohl*-V2. It should be noticed that this omission of PMs cannot be an artifact of the noisy nonstandard nature of the data. Although it is true that some writers in forums and similar discussions make highly restricted use of commas altogether, this should, under all foreseeable circumstances, affect *weil* as much as *obwohl*. However, we have verified using robust methods such as cross-validation and

25 In the corresponding GLM (Section 3.2.4, Table 6, Figure 2), the estimated coefficient for *Left=Word* is close to 0, which means that there is no effect. This is an artifact of the dummy coding of the nominal variables and the GLM intercept. The estimate $\beta_{\text{LeftWord}} = -0.019$ has to be interpreted relative to the intercept, which includes *Left=Comma*. Thus, although there is almost no evidence pointing toward differences between *Left=Comma* and *Left=Word*, *Left=Word* is virtually as different from the other levels (such as *Left=End* or *Left=Emo*) as is *Left=Comma*.

bootstrapping that there are *systematic* differences between *obwohl* and *weil* in this regard. We conclude that *obwohl* in V2 clauses is much closer to being a prototypical discourse marker than *weil* in V2 clauses. This fits in with diachronic accounts proposing that both particles are undergoing a historic development and that *obwohl* has advanced much further than *weil* (Gohl & Günthner 1999; Günthner 2003) on this route.[26] Indirectly, this should also correlate with the higher frequency of intonational pauses after *obwohl* compared with *weil* reported by Breindl (2009). Our study can, of course, not answer any questions as to what causes *obwohl* and *weil* to behave differently. By way of abduction, we simply propose that the reason why writers use different punctuation in *obwohl*-V2 and *weil*-V2 is the different grammatical—and most likely also functional—status of the constructions.

In a larger context, we consider our study to make two important contributions. First, it is yet another convincing demonstration that insisting on the discreteness of linguistic categories and strictly categorical distinctions such as *integrated* versus *nonintegrated* (see Section 2.1.1) is inadequate in the face of usage data. Writers obviously make use of a complicated network of similarity relations (in the sense of Prototype Theory and similar theories; see Section 2.2.2) when categorizing items and constructions, and this is also reflected in their writing behavior. Second, we have shown that purely writer-oriented usage-based graphemics is able to produce valuable insight into grammatical

26 Notice, however, that we have found no evidence that either of the V2 variants is more established than the other. In fact, Section 3.1 and Table 2 show that *obwohl*-V2 and *weil*-V2 occur with almost equal frequency.

phenomena by discovering systematic patterns of variation between different constructions. Because of their size and the high amount of variation contained within them, web corpora such as DECOW12Q are a valuable source of data for the task.

In our own future work, we intend to use the general methodology of mining for alternations in writers' behavior introduced here to deepen the knowledge about the interplay between syntax and graphemics and solve the many open questions in German usage-based graphemics. With the use of corpora of written language containing high amounts of spontaneously produced texts, linguists can apply a host of popular statistical techniques as used in corpus linguistics to clarify such questions—from simple non-parametric tests such as the Fisher exact test to more advanced multifactorial methods such as generalized linear modeling (for introductions to these methods, see Baayen 2008 or Gries 2013) to methods designed specifically for corpus linguistics, such as collostructional analysis (Stefanowitsch & Gries 2003).

References

Antomo, Mailin & Markus Steinbach (2010). Desintegration und Interpretation: Weil-V2-Sätze an der Schnittstelle zwischen Syntax, Semantik und Pragmatik. *Zeitschrift für Sprachwissenschaft* 29(1): 1–37.

Antomo, Mailin & Markus Steinbach (2013). Zur Semantik von Konzessivsätzen mit "obwohl." *Linguistische Berichte* 236: 427–453.

Auer, Peter & Susanne Günthner (2005). Die Entstehung von Diskursmarkern im Deutschen—ein Fall von Grammatikalisierung. In Torsten Leuschner & Tanja

Mortelsmans (eds.), *Grammatikalisierung im Deutschen*, 335–362. Berlin, New York: De Gruyter.

Augst, Gerhard, Karl Blüml, Dieter Nerius & Horst Sitta (eds.) (1997). *Zur Neuregelung der deutschen Orthographie. Begründung und Kritik*. Tübingen: Niemeyer.

Baayen, R. Harald (2008). Analyzing linguistic data. A practical introduction to statistics using R. Cambridge: Cambridge University Press.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation* 43(3): 209–226.

Barth, Danielle & Vsevolod Kapatsinski (2014, ahead of print). A multimodel inference approach to categorical variant choice: construction, priming and frequency effects on the choice between full and contracted forms of “am,” “are” and “is.” *Corpus Linguistics and Linguistic Theory*. doi:10.1515/cllt-2014-0022

Baudusch, Renate (1997). Zur Reform der Zeichensetzung—Begründung und Kommentar. In Gerhard Augst, Karl Blüml, Dieter Nerius & Horst Sitta (eds.), *Zur Neuregelung der deutschen Orthographie. Begründung und Kritik*, 243–258. Tübingen: Niemeyer.

Behrens, Ulrike (1989). Wenn nicht alle Zeichen trügen. Interpunktion als Markierung syntaktischer Konstruktionen. Frankfurt am Main: Lang.

Biemann, Chris, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski & Torsten Zesch (2013). Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics* 28(2): 23–60.

Blühdorn, Hardarik (2008). Epistemische Lesarten von Satzkonnektoren – wie sie zustande kommen und wie man sie erkennt. In Inge Pohl (ed.), *Semantik und Pragmatik—Schnittstellen*, 271–251. Frankfurt am Main: Lang.

Bredel, Ursula (2008). Die Interpunktion des Deutschen. Ein kompositionelles System zur Online-Steuerung des Lesens. Tübingen: Niemeyer.

Bredel, Ursula (2011). *Interpunktion*. Heidelberg: Winter.

Breindl, Eva (2009). Fehler mit System und Fehler im System. Topologische Varianten bei Konnektoren. In Marek Konopka & Bruno Strecker (eds.), *Deutsche Grammatik—Regeln, Normen, Sprachgebrauch. Jahrbuch des IDS Mannheim 2008*, 274–308. Berlin, New York: De Gruyter.

Bresnan, Joan (2007). Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base*, 77–96. Berlin, New York: De Gruyter.

Burnham, Kenneth P. & David R. Anderson (2002). Model selection and multimodel inference: A practical information-theoretic approach, 2nd edition. New York: Springer.

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.

Canty, Angelo & Brian Ripley (2013). *Boot: Bootstrap R (S-Plus) functions*. R package version 1.3-9. http://cran.r-project.org/package=boot

Croft, William (2001). *Radical construction grammar*. Oxford: Oxford University Press.

Croft, William (2004). Syntactic theories and syntactic methodology: A reply to Seuren. *Journal of Linguistics* 40: 637–654.

Davison, Anthony Chr. & David V. Hinkley (1997). *Bootstrap methods and their applications*. Oxford: Oxford University Press.

Divjak, Dagmar & Antti Arppe (2013). Extracting prototypes from exemplars: What can corpus data tell us about concept representation? *Cognitive Linguistics* 24(2): 221–274.

Dobrić, Nikola (2015). Three-factor prototypicality evaluation and the verb "look." *Language Sciences* 50: 1–11.

Dürscheid, Christa (2006). *Einführung in die Schriftlinguistik*, 3rd edition. Göttingen: Vandenhoek & Ruprecht.

Fabricius-Hansen, Cathrine (2011). Was wird verknüpft, mit welchen Mitteln – und wozu? Zur Mehrdimensionalität der Satzverknüpfung. In Eva Breindl, Gisella

Ferraresi & Anna Volodina (eds.), *Satzverknüpfungen. Zur Interaktion von Form, Bedeutung und Diskursfunktion*, 15–40. Berlin, New York: De Gruyter.

Fahrländer, Sarah (2013). Zur Syntax und Semantik der konzessiven Satzkonnektoren "obwohl" und "trotzdem." Mannheim: IDS.

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang & Brian Marx (2013). *Regression—models, methods, and applications*. Berlin: Springer.

Fox, John & Georges Monette (1992). Generalized collinearity diagnostics. *Journal of the American Statistics Association* 87: 178–183.

Fox, John & Sanford Weisberg (2011). *An R companion to applied regression*, 2nd edition. Thousand Oaks: Sage.

Frazier, Lyn & Keith Rayner (1988). Parametrizing language processing system: Left vs. right-branching within across languages. In John A. Hawkins (ed.), *Explaining language universals*, 247–279. Oxford: Blackwell.

Freywald, Ulrike (2008). Zur Syntax und Funktion von dass-Sätzen mit V2-Stellung. *Deutsche Sprache* 36: 246–285.

Freywald, Ulrike (2010). Obwohl vielleicht war es ganz anders. Vorüberlegungen zum Alter der Verbzweitstellung nach subordinierenden Konjunktionen. In Arne Ziegler (ed.), *Historische Textgrammatik und Historische Syntax des Deutschen*, 45–84. Berlin, New York: De Gruyter.

Freywald, Ulrike (2016). V2-Nebensätze—ein eigener Satztyp. In Rita Finkbeiner & Jörg Meibauer (eds.), *Satztypen und Konstruktionen*, 326–373. Berlin, New York: De Gruyter.

Gallmann, Peter (1996). Interpunktion (Syngrapheme). In Hartmut Günther & Otto Ludwig (eds.), *Writing and its use. An interdisciplinary handbook of international research*, vol. 2, 1456–1467. Berlin, New York: De Gruyter.

Gaumann, Ulrike (1983). "Weil die machen jetzt bald zu." Angabe- und Junktivsatz in der deutschen Gegenwartssprache. Göppingen: Göppinger Arbeiten zur Germanistik.

Gohl, Christine & Susanne Günthner (1999). Grammatikalisierung von "weil" als Diskursmarker in der gesprochenen Sprache. *Zeitschrift für Sprachwissenschaft* 18(1): 39–75.

Gries, Stefan Th. (2003). Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1: 1–27.

Gries, Stefan Th. (2013). *Statistics for linguistics with R: A practical introduction*, 2nd edition. Berlin, New York: De Gruyter.

Günthner, Susanne (1993). "... weil—man kann es ja wissenschaftlich untersuchen"—Diskurspragmatische Aspekte der Wortstellung in WEIL-Sätzen. *Linguistische Berichte* 143: 37–59.

Günthner, Susanne (1996). From subordination to coordination? Verb-second positions in German causal and concessive constructions. *Pragmatics* 6: 323–356.

Günthner, Susanne (2000). From concessive connector to discourse marker: The use of *obwohl* in everyday German interaction. In Elizabeth Couper-Kuhlen & Bernd Kortmann (eds.), *Cause—condition—concession—contrast. Cognitive discourse perspectives*, 439–468. Berlin, New York: De Gruyter.

Günthner, Susanne (2003). Lexical-grammatical variation and development. The use of conjunction as discourse markers in everyday spoken German. In Regine Eckhardt, Klaus von Heusinger & Christoph Schwarze (eds.), *Words in time. Diachronic semantics from different points of view*, 375–403. Berlin, New York: De Gruyter.

Hay, Jennifer B. & R. Harald Baayen (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Linguistics* 9(7): 342–348.

Hintzman, Douglas, L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review* 93(4): 411–428.

Holler, Anke (2009). Informationsreliefs in komplexen Sätzen: Eine diskursrelationale Analyse. *Linguistische Berichte Sonderheft* 16: 135–158.

Imo, Wolfgang (2012). Wortart Diskursmarker? In Björn Rothstein (ed.), *Nicht-flektierende Wortarten*, 48–88. Berlin, New York: De Gruyter.

Johnson, Keith (2008). *Quantitative methods in linguistics*. Hoboken: Wiley-Blackwell.

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

Kirchhoff, Frank & Beatrice Primus (2014). The architecture of punctuation systems: a historical case study of the comma in German. *Written Language and Literacy* 17(2): 195–224.

Kuperman, Victor & Joan Bresnan (2012). The effects of construction probability on word durations during spontaneous incremental sentence production. *Journal of Memory and Language* 66: 588–611.

Langacker, Ronald W. (2000). A dynamic usage-based model. In Michael Barlow & Suzanne Kemmer (eds.), *Usage-based models of language*, 1–63. Stanford: CSLI.

Langacker, Ronald W. (2008). *Cognitive grammar. A basic introduction*. Oxford: Oxford University Press.

Levine, Robert D. & Thomas E. Hukari (2006). *The unity of unbounded dependency constructions*. Stanford: CSLI.

Manning, Christopher D. (2003). Probabilistic syntax. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 289–342. Cambridge: MIT Press.

Mazuka, Reiko & Barbara Lust (1990). On parameter setting and parsing: predictions for cross-linguistic differences in adult and child processing. In Lyn Frazier & Jill de Villiers (eds.), *Language processing and language acquisition*, 163–206. Dordrecht: Kluwer.

Medin, Douglas L. & Marguerite M. Schaffer (1978). Context theory of classification learning. *Psychological Review* 85(3): 207–238.

Mentrup, Wolfgang (1983). Zur Zeichensetzung im Deutschen—Die Regeln und ihre Reform. Oder: Müssen Duden-Regeln so sein, wie sie sind? Tübingen: Narr.

Müller, Sonja (2014). Zur Anordnung der Modalpartikeln *ja* und *doch*: (In)stabile Kontexte und (non)kanonische Assertionen. *Linguistische Berichte* 238: 165–208.

Nakazawa, Minato (2014). *fmsb: Functions for medical statistics book with some demographic data*. R package version 0.4.4. http://cran.r-project.org/package=fmsb

Nerius, Dieter (ed.) (2007). *Deutsche Orthographie*. Hildesheim: Olms.

Panther, Klaus-Uwe & Klaus-Michael Köpcke (2008). A prototype approach to sentences and sentence types. *Annual Review of Cognitive Linguistics* 6(1): 83–112.

Pasch, Renate (1983). Die Kausalkonjunktionen "'da," "den" und "weil": Drei Konjunktionen—drei lexikalische Klassen. *Deutsch als Fremdsprache* 20(6): 332–337.

Pasch, Renate (1997). "Weil" mit Hauptsatz—Kuckucksei im "den"-Nest. *Deutsche Sprache* 25(3): 75–85.

Pasch, Renate, Ursula Brauße, Eva Breindl & Ulrich Hermann Waßner (2003). Handbuch der deutschen Konnektoren. Linguistische Grundlagen der Beschreibung und syntaktische Merkmale der deutschen Satzverknüpfer (Konjunktionen, Satzadverbien, Partikel). Berlin, New York: De Gruyter.

Paschke, Peter (2010). Die Interpunktion des Deutschen. Ein kompositionelles System zur Online-Steuerung des Lesens (review). *InfoDAF* 37(2/3): 144–148.

Pollard, Carl & Ivan A. Sag (1994). *Head-driven phrase structure grammar*. Stanford: CSLI.

Primus, Beatrice (1993). Sprachnorm und Sprachregularität: Das Komma im Deutschen. *Deutsche Sprache* 3: 244–263.

Primus, Beatrice (2010). Strukturelle Grundlagen des deutschen Schriftsystems. In Ursula Bredel, Astrid Müller & Gabriele Hinney (eds.), *Schriftsystem und Schrifterwerb*: Linguistisch—didaktisch—empirisch. Berlin, New York: De Gruyter.

R Core Team (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org/

Reich, Ingo & Marga Reis (2012). Koordination und Subordination. In Jörg Meibauer, Markus Steinbach & Hans Altmann (eds.), *Satztypen des Deutschen*, 536–569. Berlin, New York: De Gruyter.

Reis, Marga (2013). Weil-V2-Sätze und (k)ein Ende? Anmerkungen zur Analyse von Antomo & Steinbach. *Zeitschrift für Sprachwissenschaft* 32(2): 221–262.

Rosch, Eleanor (1973). Natural categories. *Cognitive Psychology* 4: 328–350.

Rosch, Eleanor (1978). Principles of categorization. In Eleanor Rosch & Barbara B. Lloyd (eds.), *Cognition and categorization*, 27–48. Hillsdale: Erlbaum.

Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson & Penny Boyes-Braem (1976). Basic objects in natural categories. *Cognitive Psychology* 8: 382–439.

Schäfer, Roland (2016 aop). Prototype-driven alternations: The case of German weak nouns. *Corpus Linguistics and Linguistic Theory*. Published online ahead of print.

Schäfer, Roland, Adrien Barbaresi & Felix Bildhauer (2013). The good, the bad, and the hazy: Design decisions in web corpus construction. In Stefan Evert, Egon Stemle & Paul Rayson (eds.), *Proceedings of the 8th Web as Corpus Workshop (WAC-8)*, 7–15. Lancaster: SIGWAC.

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*, 486–493. Istanbul: ELRA.

Schäfer, Roland & Felix Bildhauer (2013). *Web corpus construction*. San Francisco: Morgan & Claypool.

Schäfer, Roland & Ulrike Sayatz (2014). Die Kurzformen des Indefinitartikels im Deutschen. *Zeitschrift für Sprachwissenschaft* 33(2): 215–250.

Schwitalla, Johannes (2012). *Gesprochenes Deutsch*, 4th edition. Berlin: Schmidt.

Stefanowitsch, Anatol & Stefan T. Gries (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8(2): 209–243.

Uehara, Satoshi (2003). A diachronic perspective on prototypicality: the case of nominal adjectives in Japanese. In Hubert Cuyckens, René Dirven & John Taylor (eds.), *Cognitive approaches to lexical semantics*, 363–391. Berlin, New York: De Gruyter.

Uhmann, Susanne (1998). Verbstellungsvariation in weil-Sätzen: Lexikalische Differenzierung mit grammatischen Folgen. *Zeitschrift für Sprachwissenschaft* 17(1): 92–139.

Van Goethem, Kristel & Philippe Hiligsmann (2014). When two paths converge: debonding and clipping of Dutch "reuze." *Journal of Germanic Linguistics* 26(1): 31–64.

Volodina, Anna (2011). Konditionalität und Kausalität im Diskurs. Eine korpuslinguistische Studie zum Einfluss von Syntax und Prosodie auf die Interpretation komplexer Äußerungen. Tübingen: Narr.

Wegener, Heide (2000). Da, denn und weil—der Kampf der Konjunktionen. Zur Grammatikalisierung im kausalen Bereich. In Rolf Thieroff, Matthias Tamrat, Nanna Fuhrhop & Oliver Teuber (eds.), *Deutsche Grammatik in Theorie und Praxis*, 69–81. Tübingen: Niemeyer.

Zeschel, Arne (2008). Introduction: Usage-based approaches to language representation and processing. *Cognitive Linguistics* 19(3): 1–7.

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: Springer.

Appendix

URLs of the examples from DECOW12Q

(1a) http://www.animania.de/forum/archive/index.php/t-355.html

(1b) http://forum.animemanga.de/archive/index.php/t-4090.html

(2a) http://beautyjunkies.inbeauty.de/forum/archive/index.php/t-70474.html

(2b) http://www.kein-dsl.de/forum/archive/index.php/t-12576.html

(3a) http://s141091397.online.de/2008/10/19/2-blogeintrage-herbschden-im-weinberg/

(3b) http://www.wege-zum-pferd.de/forum/archive/index.php?t-8396.html

(4a) http://www.ipod-forum.de/ipod/ipod-allgemein/150-bilder-organisieren/

(4b) http://altes-tagebuch.mimimueller.de/ganz_alte_beitraege/

(9a) http://www.chemieonline.de/forum/archive/index.php/t-70410.html

(9b) http://www.2jesus.de/bibel-forum/turiner-grabtuch-t3971-10.html

(9c) http://www.planetheavymetal.de/drucker-3417.html

(9d) http://www.religion-studieren.de/faq/17240.html

(10) http://beautyjunkies.inbeauty.de/forum/archive/index.php/t-23192-p-311.html