Probabilistic German Morphosyntax

HABILITATIONSSCHRIFT
zur Erlangung der Lehrbefähigung für das Fach
Germanistische und Allgemeine Sprachwissenschaft

vorgelegt der Philosophischen Fakultät II
der Humboldt-Universität zu Berlin

von
Dr. Roland Schäfer
geb. am 06. Janur 1974 in Düsseldorf

Präsidentin der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekanin:
Prof. Dr. Ulrike Vedder

Berlin, den 29. Mai 2018

Gutachterinnen/Gutachter:

1. Prof. Dr. Anke Lüdeling (Humboldt-Universität zu Berlin)
2. Prof. Dr. Stefan Müller (Humboldt-Universität zu Berlin)
3. Prof. Dr. Matthias Hüning (Freie Universität Berlin)

Erklärung über Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen und den eigenen Anteil an der vorgelegten Leistung

Von den vier Zeitschriftenartikeln, die in der hier vorgelegten kumulativen Habilitationsschrift zusammengefasst wurden, entstanden die folgenden beiden in Zusammenarbeit mit Frau Dr. Ulrike Sayatz (Deutsche und niederländische Philologie, Freie Universität Berlin):

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

Ich erkläre hiermit für die beiden oben genannten Artikel gleichermaßen:

1. Die Aufarbeitung der bestehenden Literatur erfolgte zu gleichen Teilen durch Frau Dr. Sayatz und mich.
2. Die theoretische Einordnung und die Hypothesenbildung erfolgte in Diskussionen mit Frau Dr. Sayatz. Das schriftliche Ergebnis dieser Diskussionen geht zu ungefähr einem von drei Teilen auf Frau Dr. Sayatz und zu zwei von drei Teilen auf mich zurück.
3. Die empirische Forschung wurde vollumfänglich und eigenverantwortlich von mir durchgeführt.
4. Die Niederschrift erfolgte vollumfänglich durch mich.

Alle anderen Teile der vorgelegten Habilitationsschrift sind vollumfänglich durch meine eigene Leistung entstanden.

Berlin, den 29. Mai 2018

Dr. Roland Schäfer

Contents

Attachments (published papers)

Acknowledgements

First and foremost, I want to thank Ulrike Sayatz, without whom the work presented here would not exist. Among other things, Ulrike convinced me that graphemics is a highly relevant discipline and an integral part of linguistics and grammar. Together, we developed the ideas of usage-based graphemics, which underlie some of the work presented here. Ulrike also helped me to understand German morphosyntax much better through many discussions.

Furthermore, I am grateful to Felix Bildhauer for our joint work on the COW corpora (the data base underlying all of my research presented here) as well as for ongoing discussions about corpus construction, corpus analysis, and statistics (worth a substantial amount of Złoty).

I also thank Elizabeth Pankratz for being an inspiring collaborator, co-author, and moral supporter. (Not to mention proofreader under difficult circumstances.)

Moreover, I want to express my highest gratitude to (in alphabetical order) Matthias Hüning, Anke Lüdeling, and Stefan Müller for agreeing to act as referees in the official process of my *Habilitation* at *Humboldt-Universität zu Berlin*.

For their support as teachers, colleagues, employers, enablers, thesis advisors, friends, moral supporters – or any combination thereof – throughout my career, I am indebted to (in alphabetical order) Dirk Buschbom, Susanne Flach, Thomas M. Groß, Iris Hasselberg, Götz Keydana, Michael Job, Stefan Müller, Bjarne Ørsnes, Erich Poppe, Manfred Sailer, Nicolai Sinn, and Gert Webelhuth.

Student assistants who contributed significantly to the success of my research by doing a lot of painstaking work on corpus data and supervising experiments are (in alphabetical order) Sarah Dietzfelbinger, Lea Helmers, Kim Maser, and Luise Rißmann.

Finally, I want to thank everyone who supported me on a personal level (and put up with my quirks and mannerisms) during my life in the academia so far, including but surely not restricted to (in alphabetical order) Matthias B. Döring, Michael Karg, Tanja Hagedorn, my parents, and Julia Schmidt.

Preface

> Thus, [what we do] may, because of the neglect of other important structural properties, be to classify natural language along an ultimately irrelevant dimension. (Partee, ter Meulen & Wall 1990: 436–437)

I used the same quote from Partee, ter Meulen, and Wall's introduction to *mathematical methods in linguistics* in the preface to my doctoral dissertation (Schäfer 2010). In their book, the sentence alerts readers that the Chomsky hierarchy and the theory of automata might not be an adequate framework for the description of human language. In my dissertation, I used it to articulate my doubts that predicate logic (with event ontologies) and lambda calculus are adequate frameworks for the description of linguistic meaning, advocating a purely set theoretic description of sentence meaning. In the present context, I reuse the quote to mark two aspects of linguistic theory which seem important to me at the present point in the field's development.

First, in recent decades, evidence has been collected which points to the fact that language is more of a probabilistic phenomenon (where rule application is a random process governed by chance and weighted lexical, grammatical, and contextual factors) than linguists thought before the 1990s (see, for example, the programmatic paper by Bresnan 2007). Prominently, among the influencing factors are even item-specific frequency-driven effects such as the co-occurrence affinities of words to each other (as examined in the much older tradition of collocation research; Evert 2008) as well as co-occurrence affinities of words and constructions (Stefanowitsch & Gries 2003; Gries & Stefanowitsch 2004; Gries 2015a). This begs the questions of whether and how language users do not only learn rules (or *generalisations*, to use a more neutral term) but also probability distributions over rules/ generalisations and lexical items, i. e. whether the probabilistic nature of language as used is part of the linguistic knowledge or can be traced to performance effects. Since the work presented here consists of explorations in probabilistic German morphosyntax, I use the quote above to mark my belief that traditional grammatical modelling might be deficient in some relevant way. Section 1 discusses this in some detail, taking a careful stance and avoiding far-reaching claims about the architecture of the human language faculty.

Second, I find it particularly interesting that Partee, ter Meulen & Wall (1990) was called an introduction to *mathematical methods in linguistics*, and that many (but by no means all) present-day readers would expect something completely different under this label. In the textbook, the relevant mathematical methods are considered to be theoretical algebra, set theory, systems of logic, the theory of automata, etc., while many of today's readers might expect statistics from such a volume. At the time, quantitative statistical methods were not widely used in linguistics, except maybe in experimental psycholinguistics and some strains of sociolinguistics.[1] Grammar (comprising at least phonology, morphology, syntax, and referential semantics) was not seen as requiring a stochastic approach, and statistics was not part of most linguistic curricula. Thus, by taking the quote out of its original context, I want to highlight the fact that statistical analysis and statistical modelling might now be on their way to becoming *mathematical methods in linguistics* which are just as important as algebra, set theory, and the theory of automata. Ideally, statistical modelling should eventually go far beyond the use of statistics in the analysis of results obtained from corpus studies and psycholinguistic experiments, leading to integrated stochastic models of language which require knowledge of all kinds of *mathematical methods* (for example, Bod 2006). While linguistics as a discipline is clearly on its way to such an approach, a lot more theoretical and empirical work is still required.

The work presented here is theory-driven but mainly empirical. In this work, I use the methods of probabilistic grammar, specifically the now-standard methods of alternation modelling. While a lot of work exists on English alternations, German can be said to be under-researched in this kind of alternation modelling. This is surprising considering the fact that German is famous for its numerous so-called grammatical *Zweifelsfälle* 'cases of doubt' (Klein 2009; Duden 2011), which are nothing but alternations between equally acceptable forms and constructions. These phenomena are ideal test cases for probabilistic approaches.

I also present some work in which I contribute to gauging the importance and the relation between corpus data (my main source of data) and psycholinguistic experiments. Furthermore, my work makes methodological contributions by advancing relatively new sources of data (mainly web corpora), analysing non-standard language, and using and evaluating state-of-the-art statistical methods. Section 2 deals with the methodological issues in detail. While no strict formal systems have been established for the modelling of the observed effects, my work contributes to defining and delimiting the requirements to be met by future integrated formal systems

[1] Also, in some functional/cognitive circles, small-scale corpus studies were analysed using simple statistics for counts since the 1980s; see Gries (2017b: 8).

of language as represented in the minds of language users. I consider it of great importance to gather data in a methodologically sound way – as opposed to rushing linguistic theory towards another battle of frameworks (see Section 1).

There is one final point I would like to make right at the outset. My research on German was mostly published in international journals (such as Corpus Linguistics and Linguistic Theory [CLLT] and Cognitive Linguistics [COGL]). The international corpus linguistics scene is very active, with at least three major journals (International Journal of Corpus Linguistics [IJCL], Corpora, CLLT) publishing large numbers of papers per year. In 2016, in preparation for Schäfer (n.d.), an open-access introduction to statistical inference and statistical modelling for linguists, I performed a manual annotation of all 198 papers published in IJCL, CLLT, and Corpora between 2010 and 2015.[2] The list of languages covered, the corpora used, and the statistical methods used in each paper were annotated. Figure 1 shows the distribution of languages (as raw counts).[3]

English featured prominently in 136 papers (146 if World Englishes and English sign languages are added), followed by Spanish with eleven mentions as a distant second. German, on the other hand, was a major object of study in only seven papers (four in IJCL, two in Corpora, one in CLLT). Of course, this does not mean that linguists working on German (or Spanish, Chinese, Dutch, French, etc. for that matter) do not use corpora or do not publish their research. However, this result shows how corpus linguistics as a field is still very much identified with English corpus linguistics (or even BNC linguistics, see Section 2.2, esp. Figure 2 on p. 28). While this state of affairs is not detrimental for corpus linguistics, I suggest that corpus linguists working on languages other than English could benefit from taking part in the active theoretical and methodological discussions taking place in international journals. From my own point of view as a linguist working on German, it seems evident that the German language and German linguistics has a lot to contribute to current debates in corpus linguistics, especially given that German is famous for the probabilistic phenomena labelled *cases of doubt*. Thus, I hope my work encourages other linguists working on German (and other under-represented languages) to increase the visibility of their object of study in international corpus linguistics for mutual benefit. While the case studies focus strongly on German grammar,

[2] The raw data will be published with the book.

[3] Since some papers deal with more than one language, 236 language codes were assigned in total.

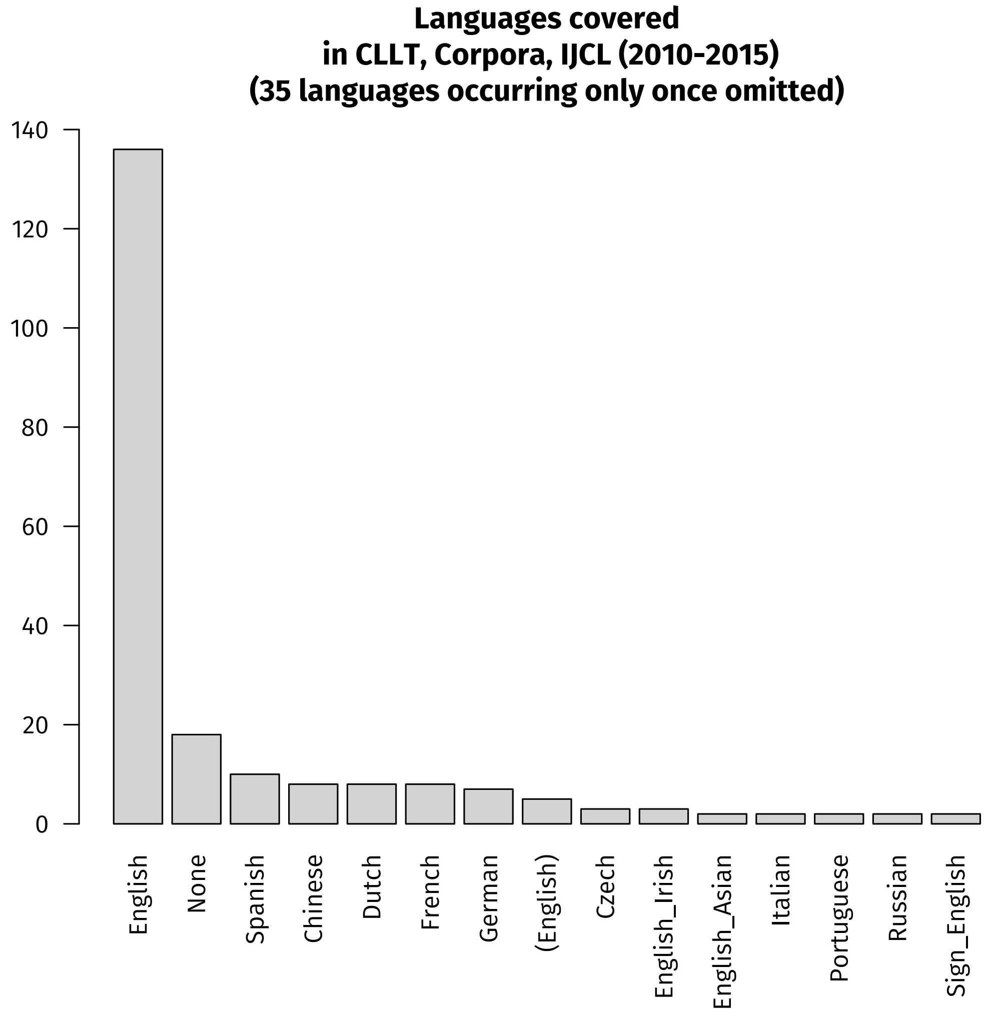

Figure 1: Languages covered in the three major corpus linguistics journals; *None* was assigned for papers which only address general or theoretical issues without reporting any original empirical work; (*English*) was assigned to papers where English is used for comparison in papers predominantly about other languages; 35 languages which only occurred once are not shown.

this general introduction predominantly takes up the foundational theoretical and methodological issues.

1 Probabilistic grammar

All case studies presented here are empirical explorations of alternation phenomena in the broad sense. While the term *alternation* is sometimes reserved for syntactic (i. e. constructional) alternations (Gries 2017b), the four case studies deal with a morphographemic alternation in the context of the development of a new paradigm of the German indefinite article (Schäfer & Sayatz 2014), a morphosyntactic alternation of so-called weak nouns which are gradually shifting towards another declension paradigm (Schäfer 2016c), a syntactic alternation between measure noun constructions (Schäfer 2018), and a phenomenon at the syntax-graphemics interface where non-standard punctuation is an obvious indicator of different clausal connections (Schäfer & Sayatz 2016).

The nature of an alternation as understood here is that language users have different forms, constructions, or even paradigms at their disposal in a given situation of language production, and that they always make a choice (unless, of course, they decide not to make the utterance).[4] While in many cases, prosodic, syntactic, lexical, pragmatic, contextual, and other constraints can be found which account for why speakers tend to choose one form or the other, these constraints appear to be soft and to interact in a weighted fashion, and there often seems to be residual free variation in speakers' choices. The observable phenomenon is thus clearly probabilistic or stochastic (as opposed to deterministic), and researchers have for a long time acknowledged this fact both for morphological phenomena (see the early overview in Hay & Baayen 2005) and (morpho)syntactic phenomena (see early contributions such as Gries 2003; Wulff 2003; Bresnan 2007).[5]

While my research clearly stands in the tradition of probabilistic grammar, I want to voice some concerns about the epistemological status of the evidence which we are gathering. In usage-based and constructionist set-

[4] For the present purpose, I understand *utterance* as comprising events of language production in both the spoken and the written mode.

[5] While the number of studies which have produced empirical evidence for the probabilistic nature of morphology and (morpho)syntax is growing, it should be noted that graphemics is under-researched in this paradigm. Writing is often not viewed as part of grammar or linguistics, and those who do advanced research on writing often view it under an acquisition perspective. Since I cannot see why phonology and phonetics – dealing with utterances realised in the spoken medium – should be treated as part of grammar but graphemics – dealing with utterances realised in the written medium – should not (see also Schäfer 2016b: 495–500), I extend the probabilistic view to graphemics. The framework of *usage-based graphemics* was therefore developed by Ulrike Sayatz and me in Schäfer & Sayatz (2016).

tings, the type of evidence as found in the papers presented here is often taken as supporting a model of grammar that does without the Chomskyan separation of competence and performance (Chomsky 1965; an overview can be found in Müller 2018: 507–518) and/or does not embrace an algebraic Aristotelian theory of language in terms of discrete linguistic categories (e. g. Manning 2002; Bod 2006; see also Kapatsinski 2014 for a recent and subjective overview in the same vein). First of all, doing away with performance altogether is clearly not a reasonable approach considering the body of psycholinguistic research showing how processing constraints affect speakers' and hearers' language use depending on factors clearly not related to learned generalisations (be they stochastic or not). As Pullum (2013a: 532) puts it,

> no sensible grammarian wants or expects grammars to yield direct representations of the raw reality of human linguistic behaviour with all its flubs, false starts and lost trains of thought.

The core question is rather whether linguistic competence itself is a probabilistic (non-Aristotelian) system (as claimed programmatically by Bresnan 2007) or whether probabilistic effects arise from performance alone.Another important and only partially related question is whether linguistic knowledge is highly specified, isolated from other forms of knowledge (including linguistic semantic knowledge as distinct from encyclopedic knowledge), and possibly also modular-serial (Fodor 1995), or linguistic knowledge (including semantics) is connected to and (mostly) an indistinguishable part of non-linguistic knowledge (see Elman 2009). While I personally favour a non-Aristotelian view which does not arbitrarily ascribe phenomena to performance, nothing in the data presented in the probabilistic tradition (which includes alternation research) conclusively forces us to assume any specific architecture as is sometimes assumed in usage-based circles Bybee & Beckner 2009. The traditional division of labour between competence and performance, together with a model-theoretic algebraic theory of syntax including an appropriate probabilistic (constraint-weighting) component, could in principle model all observed effects, including graded acceptability, stochastic alternations, and context-driven effects (e. g. Pullum 2013b: 504–507, Pullum 2013a, and Müller 2018: 499–500,507–518).[6] While some frameworks might make it (apparently) easier to model stochastic effects

[6] It must be noted, however, that while proponents of model-theoretic syntax often suggest that a probabilistic version is possible (Müller 2018: 500) or even trivial, only few (and

(sometimes at the cost of coverage or rigidity of formalisation), evidence which decides between theory A – which assumes highly specific deterministic linguistic knowledge in combination with performance effects and separate from contextual encyclopedic world knowledge – and theory B – which favours a stochastic version of linguistic knowledge, not as cleanly separated from encyclopedic knowledge – is hard to come by.

Interestingly, Elman (2009) (who is on the very far non-Aristotelian end of the Aristotelian vs. non-Aristotelian continuum) proposes a radical connectionist model (based on ample experimental evidence and computational models) which does away with the mental lexicon, a component which is in some way, shape, or form part of virtually any linguistic theory in the narrow sense (including stochastic, usage-based, constructionist theories). In simple terms, Elman proposes a model where words do not have semantic content but merely serve as cues to conceptual and world knowledge. Despite his comprehensive and erudite argumentation, he admits that the evidence and the computational models are in no way conclusive evidence for his approach (Elman 2009: 573–574). Elman (2009: 573) states:

> However, theories can also be evaluated for their ability to offer new ways of thinking about old problems, or to provoke new questions that would not be otherwise asked. A theory might be preferred over another because it leads to a research program that is more productive than the alternative.

This statement reminds one of the Kuhnian view of *normal science* as a state where a research programme generates enough new and exciting *puzzles* for researchers to solve in order to keep the field alive (Kuhn 1970). As a matter of fact, the research on alternations and other stochastic phenomena has thrived in frameworks (such as cognitive, usage-based, constructionist linguistics) which try do away (as much as possible) with the competence-performance dichotomy and with a highly modular and specific model of linguistic competence and *not* in generative frameworks (such as Chomskyan minimalism) or model-theoretic frameworks (such as Head-Driven Phrase Structure Grammar), which is why the term *probabilistic linguistics* has become associated with the former type of framework. My research is therefore presented with reference to usage-based approaches, prominently

often sketchy and inconsequential) attempts have been made to deliver actual implementations (e. g. Arnold & Linardaki 2007).

addressing the prototype vs. exemplar debate.[7] In a spirit similar to the Jeffrey Elman quote above, I consider the usage-based framework and the associated community the one which currently *offers new ways of thinking about old problems, and which provokes new questions that would not be otherwise asked*, or, in Kuhnian terms, which *offers enough new puzzles to solve* when it comes to stochastic surface effects in language use. In no way does this mean that I consider the empirical findings intrinsically incompatible with other frameworks. As soon as people from such frameworks develop a significant interest in modelling my data, they may do so.[8]

All that said, I see my work less as speaking in favour of any specific linguistic framework and more as making theoretic and above all methodological contributions to some very specific questions, such as the prototype vs. exemplar debate (in Schäfer 2016c, Schäfer & Sayatz 2016, and Schäfer 2018), paradigmatic morphology (in Schäfer & Sayatz 2014 and Schäfer 2016c), graphemics under a usage-based perspective (in Schäfer & Sayatz 2016), the experimental cross-validation of corpus-derived models (in Schäfer 2018).[9] All case studies promote the use of web corpora as ideal sources of data, which is argued for in Section 2.2. Finally, the statistical methods used in the analysis of corpus and experimental data are one of my primary foci, which is why Section 2.3 provides a short overview of statistics and scientific inference.

[7] However, it does without a specific commitment to constructionist approaches or cognitive grammar in the narrow sense of Langacker (1987).

[8] Such attempts would be facilitated by the fact that I publish all data related to my published research freely (see https://github.com/rsling).

[9] With respect to the prototype vs. exemplar debate, however, a similar situation is described for corpus linguistics in Section 2.1. The data only provide limited cues as to which theory is more appropriate.

2 Theories, methods, and data

2.1 Prototypes and exemplars

In this section I discuss one major overarching theoretical theme of the case studies, namely prototype theory and its rival, exemplar theory. The central question is which type of cognitive representation research on alternations provides evidence for.[10] The typical approach in alternation research is to annotate a large number of corpus sentences with linguistic features and to model the probability of the variants being chosen given these features. The idea is that a variant is chosen when the influencing features cumulatively assume typical values for that variant. For several reasons, a variant of prototype theory with features (Rosch 1978) is a good candidate for the appropriate cognitive model. I will first introduce prototype theory and exemplar theory, including a discussion of the state of the art in linguistics and cognitive science. The discussion then focusses specifically on the data-driven corpus-based approaches which have been used very prominently in probabilistic modelling and alternation research (see Gries 2017b), and whether such approaches have anything to contribute to the prototype vs. exemplar debate.

In Aristotelian approaches to linguistic categorisation, category membership is determined by rules and defining features, and it is consequently not viewed as a matter of degree (Sutcliffe 1993; G. Murphy 2002: 11–16). However, based on evidence pointing to the fact that humans often categorise objects by similarity and with varying degrees of fuzziness, prototype theory (Rosch 1973; see Taylor 2008 for an overview) and exemplar theory (Medin & Schaffer 1978; Hintzman 1986) were developed. Prototype theory assumes that categories are defined by the similarity of their members to a mentally stored abstraction. This abstraction takes the form of the most prototypical member or – in later versions of the theory – weighted features defining the prototype (see Schäfer 2016c for a detailed description of prototype theory with features including *cue validity*). Many researchers such as Gries (2003); Gilquin (2006); Nesset & Janda (2010); Dobrić (2015) have used prototype theory in some form of corpus-based linguistic modelling.[11]

[10] This is an extended and modified version of Section 1 from Schäfer (2018) with additions from Schäfer (2016c).

[11] In the influential framework of cognitive grammar (Langacker 1987), prototypes (which, as should be remembered, represent abstractions already) are literally taken as prototypical exemplars, and there is an additional level of fully discrete abstractions in the form of schemas. Schemas are characterised by the properties common to all members of the

While prototype theory is well suited for modeling constructional choices, it has a prominent adversary in exemplar theory (Medin & Schaffer 1978; Hintzman 1986). Prototype theory and exemplar theory model essentially the same types of effects but differ significantly in whether they assume higher-level abstractions in the form of single maximally prototypical exemplars or their features (prototype theory) or assume that categories emerge through the storage of many exemplars and similarity classification on those exemplars (exemplar theory). Barsalou (1990) already showed that prototype and exemplar theory model the same types of surface effects and are informationally equivalent, at least when it comes to the results of cognitive agents' behaviour. Barsalou (1990: 84) states that

> we can not say whether category knowledge is distributed in exemplars or centralized abstractions. But we do know that any account of knowledge that excludes idiosyncratic information, cooccurrence information, or dynamic representation is inadequate.

Consequently, research producing evidence in favour of one theory or the other commonly does not use mere output data but tests the procedural behaviour of subjects in controlled experiments, for example the speed of category retrieval. In very early experiments, Posner & Keele (1968) showed, for example, that highly prototypical unseen exemplars were categorised more easily by subjects compared to less prototypical ones which had been included in the data made available to them in order to learn the categories. This was (at least at the time) taken as evidence that subjects categorise by prototypes. Since corpus data only show artefacts of production events and we have no experimental access to the speaker's or writer's

category, whereas a prototypical category member might have very specific additional properties not at all shared by all or even most members (Langacker 1987: 371-375). The prototype can serve as a reference point when classifying new objects which do not share all properties of the schema, but this would (if repeated) lead to the creation of an even more abstract (hierarchically higher and less specific) schema which describes the new member and the ones belonging to the previous schema. As pointed out by Langacker (1987: 136–137), schemas and prototypes thus fulfil different roles and can be assumed to co-exist. A strict exemplar view of language is incompatible, as far as I can see, with Langacker's view of schemas, but any theory of categorisation that allows for at least some kind of abstraction is not in fundamental contradiction with it. In my research, I do not use schemas in my descriptions of the relevant categories, mostly because the aspect of similarity and fuzzy classification is central to probabilistic modelling, and a formulation in terms of schemas would bring about an unnecessarily high degree of abstraction (see Taylor 2003: 70–71 for a parallel argument).

performance and their actual similarity judgements, one should be sceptical whether corpus analysis alone could ever decide which theory of mental representation is more suitable. Gries (2003: 22) can be taken as recognising this, when he says:

> Frequently, Rosch's results were […] interpreted as if they were statements on the structure of mental representations as such; cf. the effects = structure fallacy and the prototype = representation fallacy. I do not wish to support such interpretations. […] Still, even if the form of analysis does not translate into statements on mental representations, the high predictive power […] shows that the cognitive factors underlying the choice of construction have been identified properly and weighted in accordance with their importance for actual usage.

A similar caveat (without direct reference to prototypes and exemplars) can be found in Dąbrowska (2016: 486–487), who states that we cannot "deduce mental representations from patterns of use", i. e. from corpus data. As corpus data are artefacts of cognitive agents' behaviour, they cannot decide between two theories for which Barsalou's informational equivalence criterion holds.

Given this situation, the question arises of how the discussion in the usage-based linguistics community connects with the current discussion in cognitive science. In cognitive science, it is mostly accepted that exemplar theories have greater explanatory power (Vanpaemel 2016: 184) and that abstraction is only needed marginally, if at all.[12] Still, various attempts have been made over the past decades to settle the dispute between abstraction-based models (models with rules or prototypes) and exemplar models, or to find models which unite the two extremes. Vanpaemel & Storms (2008) and M. D. Lee & Vanpaemel (2008) proposed the *varying abstraction model*

[12] The hard empirical evidence in favour of exemplar models is substantial. For example, in Hahn et al. (2010), the authors show that subjects even use exemplar similarity over abstract knowledge even when they are given very simple explicit rules to be learned. This is highly relevant because most other studies focus on the learning of implicit rule-based knowledge, which involves many auxiliary assumptions in actual experiments (Hahn et al. 2010: 2). On the other hand, there is evidence that neither theory is fully adequate to model humans' capabilities to form categories. For example, Conaway & Kurtz (2016) show that both prototype theory and exemplar theory fail to explain certain experimental results where subjects learn to generalise beyond the input in a way that cannot be explained by similarity.

(VAM) which “attempt[s] to balance economy and informativeness” (M. D. Lee & Vanpaemel 2008: 745), treating models with full abstraction (radical prototype theory) and no abstraction at all (radical exemplar theory) as special cases of a model which allows for both abstraction and exemplar effects. The mixture model of categorisation (MMC) by Rosseel (2002) is a model with abstraction in the form of hierarchical clusters of exemplars, and these clusters of objects are characterised by a probability distribution over their features, and categorising new objects is a process of estimating the probability of this object belonging to one of the clusters. Griffiths et al. (2009) go further and present a computational model which is able to choose the appropriate complexity of representation for a given category. However, despite these (and more) attempts to reconcile or unite the two approaches while developing spelled-out mathematical models, Vanpaemel (2016: 183–184) describes the state of affairs between adherents of neo-prototype theory (such as Minda & Smith 2001; 2002) and exemplar theory as a stalemate.

In cognitive linguistics, Divjak & Arppe (2013) is a very rare example of a paper where such issues are taken up with reference to the current research in cognitive science. Their corpus-based approach shows “one way of systematically analyzing usage data as contained in corpora to yield a scheme, compatible with usage-based theories of language, by which the assumptions of both the prototype and exemplar theories can be operationalized” (Divjak & Arppe 2013: 267). Their approach to implementing a varying abstraction model (Divjak & Arppe 2013: 254–260) is based on hierarchical clustering of annotated properties of sentences. They cluster sentences containing Russian verbs of trying. Then, they single out the one sentence from each cluster which scores the highest probability for any of the six *try* verbs according to a polytomous regression model estimated on the same data. The clusters are interpreted as intermediate-level exemplar-derived abstractions of typical contexts for these high-probability verbs (typically more than one cluster for each verb; Divjak & Arppe 2013: 255–256). The crucial difference between such data-driven corpus-based analyses and experiments in cognitive science (Divjak & Arppe 2013 use Verbeemen et al. 2007 as their reference) is that cognitive research is based on experiments where subjects produce actual category assignments or similarity judgements, and in corpus studies, the categories and category membership are determined purely from existing data. The experimental approach with reduced and/or artificial stimuli makes it much easier to examine very specific effects in the behaviour of the subjects. While I do not think the results of the case study presented in Divjak & Arppe (2013) are invalid, any data set can be analysed to yield a certain number of clusters, and this fact alone does not substantiate any claim about one mental representation or another. Thus, the study does not ensure that the clusters emerging from the data correspond to any speaker's cognitive representation. In other words, Divjak & Arppe (2013: 229–230) fall victim to Barsalou's equivalence trap when they state without further motivation that

> [t]he objectives of this study are, first, to explore how the prototype and exemplar models of categorization manifest themselves in corpus data […]. Although corpus data do not reflect the characteristics of mental grammars directly, we do consider corpus data a legitimate source of data about mental grammars.

The second sentence of this quote has at least one reading in which it is contradictory. Compare the (already mentioned) more realistic views in Barsalou (1990: 84), Dąbrowska (2016: 486–487) and Gries (2003: 22).

As mentioned above, in cognitive science, experimental setups which allow access to the cognitive agents' performance over time are preferred in order to produce evidence for either one of the two competing theories. See Storms, Boeck & Ruts (2000) for a comparison of the theories in different experimental settings. However, the trade-off one has to accept when doing experiments with highly simplified stimuli and very simple tasks is their lower *external validity* (i. e. their lower degree of generalisability) and their high dependence on potentially problematic operationalisations of constructs, control of confounding factors in the face of a limited number of available subjects, etc. (in other words, critical dependence on *construct validity* and *internal validity*).[13] Tasks in cognitive science have been criticised exactly for their lack of external validity, for example by G. L. Murphy (2003). From a linguistic perspective, it is remarkable in this context that Voorspoels, Vanpaemel & Storms (2011) consider their experimental task – which is the assignment of typicality scores to nouns from the domains of *animals* and *artefacts* to categories like *bird*, *fish*, *clothing*, or *tools* – a study of "superordinate natural language categories, whereas most

[13] Construct validity requires the measurements made in a experiment to be credible and substantive indicators of a theoretically postulated construct (such as a prototype). Under internal validity an experiment establishes a causal relationship between experimental manipulations and the measured effects through minimisation of systematic measurement error. An accessible overview of the different types of validity can be found in Chapter 1 of Maxwell & Delaney (2004). The discussion of types of validity goes back to Cronbach & Meehl (1955); Campbell & Fiske (1959).

evidence supporting exemplar representations has been found in artificial categories of a more subordinate level" (Voorspoels, Vanpaemel & Storms 2011: 1013). Corpus linguists interested in probabilistic alternation modelling deal with much more complex high-level categories and use large and complex feature sets, especially in (morpho-)syntax.[14] It is thus an advantage of much linguistic work on categorisation that it deals with complex and realistically produced data, because this greatly improves the external validity of studies, albeit by sacrificing some construct validity. An ideal contribution by cognitive corpus linguists to the research on (levels of) category abstraction in the human mind would thus be to provide analyses which have great external validity and complexity while carefully making sure that (and determining to what extent) these finding correlate with reactions from cognitive agents under more controlled experimental conditions, which increases the construct validity. This is why experimental validations of corpus-derived models should under all circumstances become the standard procedure. Section 2.2 briefly discusses this approach.

In closing, I want to point out that my work often conveniently uses prototype-theoretical formulations (Schäfer & Sayatz 2016, Schäfer 2016c, Schäfer 2018) because the high-level contentful features which are mostly used in probabilistic modelling (such as semantic classes of lemmas, definiteness of noun phrases, discourse status, or register, to name just a few) invite a description that allows for abstractions. However, Schäfer (2018) argues that certain types of effects are at least implausible to model as abstractions. In the study, it is shown that lemma frequency and construction-lemma attraction influence the choice of alternants. Such item-specific effects indeed appear to favour an exemplar view. However, it must be noted that it is always possible that item-specific effects can in fact be traced back to abstraction effects (such as semantic properties of lemmas). Again, in the spirit of Section 1, we should be aware that our inferences are more often than not abductive, i. e. inferences to the (in the view of our research community) best explanation.

[14] Notice, however, that recently, approaches have emerged which solve at least some problems by abandoning linguistic high-level features altogether (Baayen, Shaoul, et al. 2016; Ramscar & Port 2016). Clearly, they have not (or at least not yet) reached mainstream popularity, and it remains to be seen how well they perform on a broader range of questions.

2.2 Corpora in cognitively oriented linguistics

2.2.1 Problems with corpora and some solutions

The empirical analysis of probabilistic phenomena such as alternations requires researchers to make choices with regard to the data they use for their scientific inferences. Depending on the amount of data available and type of inference, methods for the numerical analysis of the data are also required. In the present section, I argue why web corpora (i. e. corpora built using material collected from the WWW) are ideal for the corpus-based work presented here, and I briefly introduce the idea of experimental validation of corpus findings. Section 2.3 will then discuss methods of statistical analysis.

Corpora have been used as a major source of data in alternation research and cognitively oriented linguistics in general, and my research is no exception. Since cognitively oriented linguistics is an attempt to model cognitive representations as well as the cognitive mechanisms involved in using these representations to produce and understand utterances, the question arises whether corpus data – i. e. artefacts of language use – are an appropriate source of data in cognitively oriented linguistics.

Prominently, Gries (2017a: 591–592) argues that corpus linguistics is essentially the quantitative analyses of co-occurrence frequencies (e. g. of words and words or words and constructions, words and senses) in collections of texts, which is often related to the *distributional hypothesis* and traced back to Harris (1954). Gries also notes that the major tenet of cognitively oriented linguistics is that language users learn language by acquiring knowledge about the probabilities of words, constructions, senses, etc. in a given context (in the broadest sense of the word *context*). Thus, Gries concludes, both disciplines deal with distributional phenomena and are highly compatible. Clearly, this is accurate inasmuch as both corpus linguistics and cognitively oriented linguistics examine types of distributional phenomena, but one distributional phenomenon is not necessarily like any other one. The implicit claim made by Gries is that both fields deal with *the same or at least two highly and causally related distributional phenomena*. While it is impossible to refute this implicit additional assumption, it is also difficult to substantiate it. Therefore, I suggest that we accept it as a working hypothesis. At least, however, the approach begs the question of whether corpora represent the cognitive reality of language users in any meaningful and reliable way.[15] The traditional discussion of the *representativeness* of a corpus

[15] I do, however, contradict the categorically contemptuous tone found in many works from cognitively oriented linguistics (Gries 2017a: 590–593 is no exception) against earlier work

does not necessarily help in this context, because it is more often than not centred around the concept of a corpus being *representative of a language* as a whole, using as points of reference: (i) the distribution of texts or text types in the output of all speakers of a language (production-based), (ii) the distribution of the relevance of texts or text types in the whole speech community (relevance-based), or (iii) the distribution of speakers' exposure to different texts or text types (perception-based).[16]

Indeed, given the argumentation from Gries (2017a) discussed above, a perception-based view of corpora seems to be the most appropriate for a cognitive approach to language where input frequencies play the most crucial role. However, the linguistic experience of language users is most definitely a highly individual matter, and most corpora force researchers to work with highly problematic abstractions. In a recent contribution where the perception-based view is argued to be valid in cognitively oriented corpus linguistics, Stefanowitsch & Flach (2016: 104) take a quite cavalier stance on representativeness:[17]

> In this wider context, large, register-mixed corpora such as the British National Corpus […] may not be perfect models of the linguistic experience of adult speakers, but they are reasonably close to the input of an idealized average member of the relevant speech community.

The assumption that an *idealised average speaker* is a valid construct seems naïve at best.[18] If the concept of an idealised average speaker were admissible and if there were indeed corpora (like the BNC) representative of this idealised speaker's input, then different actual speakers should not learn considerably diverging grammars. If it turns out that actual speakers

based on researchers' intuitions. As Sprouse & Almeida (2012); Sprouse, Schütze & Almeida (2013) have shown with significant methodological rigour, judgements collected in a traditional way through intuition by linguists can be highly reliable. The question is rather which types of data are used as evidence to support which claims. Intuitive judgements (even by linguists) are not intrinsically unreliable; they might just be the wrong tool in certain situations.

[16] For overviews from different perspectives, see Biber (1993), McEnery, Xiao & Tono (2006), Leech (2007), Hunston (2008). A summary of the discussion is found in Chapter 5 of Schäfer & Bildhauer (2013).

[17] The picture does not change significantly if corpora are seen as collections of linguistic output events under a cognitively oriented perspective (Tummers, Heylen & Geeraerts 2005).

[18] Essentially, they argue for a license to conclude that any distributional pattern found in the BNC automatically has a cognitive reality. If life were this easy, many more researchers would certainly use the BNC exclusively.

indeed acquire fundamentally different grammars, however, then the idealisation is unwarranted. There is growing evidence from both psycholinguistic research and cognitively oriented linguistics that differences between competent adult speakers of a language are substantial and should not be averaged. This concerns speakers' performance in linguistic tasks (Huettig & Janse 2016 and references therein), but it also affects their individual grammars. For example, Dąbrowska (2008; 2012) clearly found that there is no convergence of the grammars of different Polish adult speakers towards a unified grammar (with respect to the phenomena under study), and Dąbrowska (2015) puts this into a larger picture. This does, of course, not entail that speakers would perceive each other's languages as *different* like foreign languages or dialects – or that they would be expected to have problems communicating in their everyday life. However, this clearly questions the usefulness of the concept of an *idealised average speaker*. A commonly adopted solution is to model speaker-variation as a nuisance variable, usually by adding a per-speaker random effect to the statistical models (Gries 2015b; Schäfer 2018; see Section 2.3). Most corpora, however, lack metadata to identify the authors of specific texts reliably. If they do, single authors usually contribute far too few data to take individual variation into account in an informative way. But even if there are enough data, the random-effect approach is just a way of taking care of unmodelled heterogeneity (see also Section 2.3.4). It does not make the remaining *averaged* part of the model more cognitively real.[19]

In the corpus linguistic discussion on representativeness (as mentioned above), this problem has received little attention, mostly because this discussion has traditionally focussed on a global notion of representativeness. An ideal corpus of a language is assumed to be representative of a language (such as English or German) as a whole (whatever that means). This holistic approach to the question of representativeness (which surely still inspired the view argued for by Stefanowitsch & Flach 2016) is not applicable to cognitively oriented linguistics, which probably requires techniques for experiments and observational studies similar to those in the social sciences, psychology, and cognitive science. The goal in scientific experiments (or scientific studies, to use a more general term) is to make inferences about a *population of interest* using a sample of data from that population. The population of interest has to be defined with regard to each experiment individually, and it might be something very specific (such as the written

[19] Although without specific empirical backup, Newmeyer (2003: 695–698) already made this point convincingly, which is something one can admit without committing to the full argument he put forward.

output of speakers of a certain age, in a specific register, etc.) instead of *the language* or *the average speaker* (across all communicative settings and modes). In the social sciences, the concepts of *global and specific representativeness* (Bortz 2005: 86) are used to describe the relevant distinction. The crucial point is thus not whether a corpus is *representative of a language* but whether a sample taken for a specific purpose represents the population of interest for the concrete study.

A standard approach implicitly adopted in many corpus studies (including mine) is to define the population of interest as *corpus exemplars in which a certain range of constructions, words, etc. occurs*.[20] This is effectively what we do if we run unrestricted queries looking for specific morphological or syntactic patterns in some large corpus like the DeReKo of the Institut für Deutsche Sprache (IDS; Kupietz et al. 2010) or the DECOW16A (Schäfer & Bildhauer 2012; Biemann et al. 2013; Schäfer & Bildhauer 2013; Schäfer 2015). With certain caveats taken into consideration and given the right research question, nothing speaks against this approach, as I will argue below. Going beyond this unrestricted query approach, however, Gries (2015b; 2017a) argues that by using metadata (as available for the British National Corpus; D. Lee 2001; Burnard 2007), searches can be refined to specifically examine phenomena with different (relative) frequencies in different modes, genres, etc. If there is no hypothesis that mode, register, etc. have an influence on (relative) frequencies, however, such refinements are not strictly required.[21]

The unrestricted query approach works when the hypothesis is that the relative frequencies of two structures A and B change when going from condition x to condition y. Mathematically, this hypothesis can be expressed as (1) (see also Section 2.3.3).[22]

(1)
$$\frac{f(A|x)}{f(B|x)} \neq \frac{f(A|y)}{f(B|y)}$$

For example, in Schäfer (2016c) one hypothesis was that the frequency of the singular non-nominative form of a weak noun such as (*den/dem/des*) *Planeten* ‘planet’ (structure A) compared to the frequency of its (non-

[20] This implicitly assumes something similar to what Berk & Freedman (2009: 27) call an *imaginary population*. Their criticism applies only in limited ways to corpus sampling given the argument I make here.

[21] Note that the word *influence* implies a *causal relationship*.

[22] f(A|x) is to be read as in the notation for conditional probabilities, except with frequencies instead of probabilities. It is thus the *frequency of item A under condition x*.

standard) strong form (*den/dem/des*) *Planet* (structure B) is different in the genitive (condition x) than in the accusative and dative (condition y).

As long as a corpus contains both structures and *given the randomness assumption* (RA), the hypothesis can be examined using an unrestricted query. The RA holds if all other potential influencing factors which favour the occurrence of A or B are distributed equally in the conditions x and y. In the example, individual lemmas might have a tendency to favour the strong form over the weak form or vice versa, or different registers might favour one form over the other etc. However, given a random assignment of lemmas, registers, and so on, condition x and condition y would always yield a different distribution of A and B under hypothesis (1). The RA could thus still hold even if the strong (non-standard) form occurred predominantly in a specific register or mode, and if (additionally) the corpus did not contain very many texts from this register or mode.[23] However, effects might be more difficult to detect in such a situation; see Section 2.3.3 for more on this. In the worst case, a corpus might simply contain not enough exemplars of a certain phenomenon as a result of an inappropriate register or mode composition. This would be detrimental for any study, but it is also definitely not related to the RA. In any case, problems with the RA are not exclusive to linguistics or corpus linguistics, but represent standard problems with data sampling for experiments.[24]

On the basis of these elaborations, I propose that the core problem for corpus studies in cognitively oriented linguistics is not the RA. Rather, it is one of the following problems (depending on the research question):

- **Problem 1**: The corpus contains pooled output data from numerous individual speakers, making it impossible or difficult to draw conclusions about cognitive representations (see Dąbrowska 2008; 2012; 2015).
- **Problem 2**: The corpus does not contain the relevant meta-information to draw a sample which represents the population of interest in a given study (for example, if genre or register effects are of interest and the corpus does not contain the relevant metadata).

[23] In the given example, this might be the case with the DeReKo. It contains predominantly edited newspaper texts, and the strong forms – being non-standard – are probably very rare in such texts.

[24] "Drawing a random sample of the U. S. population, in this technical sense, would cost several billion dollars (since it requires a census as a preliminary matter) and would probably require the suspension of major constitutional guarantees. Random sampling is not an idea to be lightly invoked." (Berk & Freedman 2009: 23)

- **Problem 3**: Even if the composition of the corpus in terms of registers, modes, etc. is not a primary research interest, a phenomenon which occurs only in specific registers, modes, etc. might be underrepresented in a given corpus because of its composition.

Problem 1 has multiple remedies. At first sight, it might seem a valid option to enrich corpora with metadata such that the writer/speaker of each exemplar can be identified, then taking per-speaker preferences into account (as promoted by Gries 2015b; 2017b). However, if the mental grammar of each speaker varies (as suggested by Dąbrowska 2008; 2012; 2015), then it would not be sufficient to take per-speaker *tendencies* into account in the sense of: *speaker i favours variant A over variant B with probability* p_i. Instead, a separate (statistical) model (or a complex model with per-speaker random slopes) would have to be built for each speaker, since the influencing factors guiding each speaker's grammatical choices would be weighted differently. This would lead either to extremely complex over-parametrised models (if random intercepts and slopes are used; see Section 2.3) or to a lot of different models with no way to come up with an interesting generalisation. Kuperman & Bresnan (2012) suggest (in the context of an experimental setting) to use multi-model averaging (Anderson & Burnham 2002) to take into account the variation between speaker-grammars (see also Barth & Kapatsinski 2014). While this might be applicable for psycholinguistic experiments, it is clearly not feasible in corpus studies with large numbers of speakers, especially since the models over which one averages should be known theoretical options. In any case, the per-speaker data would be too sparse in any conceivable corpus to make such over-parametrised modelling feasible.

Another more realistic option is to focus on the cognitive principles that theories of cognition predict should govern the formation of mentally represented grammars. Such cognitive principles should be observable with considerable stability across groups of speakers. In Schäfer (2016c), for example, I used predictions derived from Köpcke (1995) about the prototype representations of weak nouns (in the sense of inter-individual cognitive principles) to derive predictions for the outcome of the corpus study. In Schäfer (2018), I used arguments from grammaticalisation research (Koptjevskaja-Tamm 2001) to argue for plausible general cognitive mechanisms which guide the relevant choice between two pseudo-partitive NP structures. Based on such assumptions, the use of massively pooled data from large corpora is not unjustified, and it avoids the dangers of data dredging and fishing for spurious correlations (Good & Hardin 2012). Even under the assumption of such theoretical predictions about general cognitive mechanisms, it should become standard practice to evaluate corpus-derived models using experimental techniques to check whether corpus data and reactions by native speakers under controlled conditions converge. A review of the state of the art was provided by Newman & Sorenson Duncan 2015, who enumerate a number of studies showing how corpus data and experimental data converge (such as Bresnan et al. 2007; Durrant & Doherty 2010; Gries & Wulff 2005; Gries, Hampe & Schönefeld 2005) and a number of studies where the two types of data led to diverging or only partially converging results (such as Arppe & Järvikivi 2007; Dąbrowska 2014; Mollin 2009). When researchers do not achieve convergence, they often try to explain this by differentiating between the actual cognitive construct and whatever the pooled usage data as found in corpora represent. For example, Dąbrowska (2014: 411) lists a number of possible reasons to explain why subjects in her experiment diverged in their word association preferences from collocation measures extracted from corpora. Alternatively, researchers argue for a more adequate statistical analysis to increase the fit between corpus data and experimental data. See, for example Divjak, Arppe & Baayen 2016, who show that generalised additive models (GAMs) are better suited than generalised linear models (GLMs) for correlating reading times and corpus data. No general consensus and no commonly accepted best-practice approach has emerged so far, which is not surprising given the number of cognitive constructs assumed at diverse levels, the problems of corpus composition, the operationalisations involved in experiments, and the choice of statistical tools. Also, experimental validation of corpus-based findings has simply not become a general requirement. As Divjak, Dąbrowska & Arppe (2016: 3–4) put it:

> There are now a number of published multivariate models that use data[,] extracted from corpora […] to predict the choice for one morpheme, lexeme or construction over another. However, […] only a small number of these corpus-based studies have been cross-validated […]. Of these cross-validated studies, few have directly evaluated the prediction accuracy of a complex, multivariate corpus-based model on humans using authentic corpus sentences […].

Therefore, in Schäfer (2018), I used experimental validation in two different paradigms and provide possible explanations for the quality of the fit between the corpus data and the experimental data, much in the spirit of Dąbrowska (2014).[25]

Turning to **problem 2**, there is a relatively simple solution in theory, which might, however, be difficult to implement in practice. If the corpus lacks the appropriate metadata, one can simply draw an unrestricted sample and manually annotate the relevant registers, styles, etc. afterwards. While this might sound quite laborious, it also ensures that the relevant categories are the ones the researcher has a theory-driven hypothesis about. Given the many different definitions of registers and similar categories, it is not likely that corpus creators would annotate a corpus with exactly the taxonomy the researcher has in mind.[26] Of all of my studies collected here, only Schäfer (2018) made reference to style (not register) as an influencing factor. Fortunately, as one of the creators of the DECOW corpora, I could implement the necessary technology to automatically annotate the corpus with proxy variables to style (see Schäfer 2018, also Schäfer, Barbaresi & Bildhauer 2013).

Finally, **problem 3** has a relatively simple solution. Under the unrestricted query approach (which – I want to stress once again – was used for all research collected here), data sparsity can be remedied by creating larger and at the same time more varied corpora. Research on phenomena which are simply rare in general benefit from sheer corpus size. In all case studies collected here, one of the examined variants had a low frequency, and the studies all benefitted from the fact that the corpus used (DECOW in different versions) was very large. In Schäfer (2016c), the proportion of forms of the weak nouns inflected according to the strong paradigm was reported to be between approximately 1% and 2%. In Schäfer & Sayatz (2014), the forms of the cliticised indefinite article was estimated to account for roughly 2.5% of all forms of the indefinite article. Furthermore, in Schäfer & Sayatz (2016), clauses headed by *obwohl* ‘although, then again’ and *weil* ‘because’ showing verb-second order made up roughly 6% to 7% of all clauses headed by those particles. Only in Schäfer (2018) was the situation less extreme, with the rarer of the two competing measure NP constructions accounting for approximately 22% of all exemplars in the sample.

Furthermore, DECOW is a web-derived corpus created by an unrestricted crawl of the German-speaking web. As such, it contains documents written in all sorts of styles, registers, and text types. These include sources of non-

[25] See also Schäfer & Pankratz (2018), where a different type of experimental validation of corpus-based findings is used.

[26] Also notice that recent approaches to automatic large-scale register identification failed with a classification accuracy in the region of 50% (Biber & Egbert 2016).

standard written language such as forums, which is not true of the other very large German corpus, the DeReKo. The fact that such sources are included in the corpus was vital for at least Schäfer & Sayatz (2014), Schäfer (2016c), and Schäfer & Sayatz (2016) because the relevant alternation is only truly productive in non-standard language, as normative grammars ban one of the alternants.[27] With any other available corpus, the case studies would have run into problem 3 (data sparsity due to an inappropriate composition of the corpus in terms of registers, modes, text types, etc.). In Section 2.2.2, I therefore provide a short motivation of why web corpora are an important new source of data.

2.2.2 Web corpora

Web corpora were made popular through the WaCky initiative (Baroni et al. 2009) starting around 2005. The WaCky corpora were attractive to many researchers because they were made available freely and could be downloaded fully, which allows for all kinds of local processing not possible through web-based query interfaces. In parallel, the SketchEngine corpora were developed as a commercial alternative (Kilgarriff et al. 2014), and the SketchEngine project probably represents the most significant current provider of web-derived linguistic corpora. The COW corpora have been under development since 2012 (Schäfer & Bildhauer 2012; Schäfer, Barbaresi & Bildhauer 2013; Schäfer & Bildhauer 2013; Schäfer 2015; Bildhauer & Schäfer 2016; Schäfer 2016a; 2017; Bildhauer & Schäfer 2017). Like the WaCky corpora, they are freely available both for download and via a web interface for easy querying.[28]

While smaller specialised corpora are sometimes derived from web data (e. g. Krause 2016), the major advantages of web data in the context of the research presented here (see especially problem 3 from Section 2.2.1) is that there is a virtually unlimited supply of textual data available on the web. Also, web data includes non-standard written forms, and the breadth of the variation contained within it is enormous. For example, in Bildhauer

[27] Some normative grammars take similarly clear stances on the phenomenon discussed in Schäfer (2018), as pointed out in the paper. However, the normatively dispreferred variant is still used quite often, despite such attempts to suppress it.

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

& Schäfer (2016; 2017), it was shown that the DECOW16 web corpus has a much broader spread of topics than the DeReKo newspaper corpus.[29]

Thus, web corpora were the obvious choice for the case studies, and the DECOW corpus was created by me specifically for the purpose of conducting my linguistic research published between 2014 and today.[30] Given the data from the survey to be published in Schäfer (n.d.) (see p. 5), we can assess the impact web corpora have had on current research in corpus linguistics by looking at the frequency with which different corpora have been used in research published in the three major international corpus linguistics journals between 2010 and 2015. Figure 2 plots the distribution of the corpora used (328 usages of corpora in 198 papers).

Figure 2 clearly shows that the distribution of the use of corpora follows almost a power law distribution. Custom corpora built for a specific research project are most frequent (used 68 times), followed by the BNC (used 36 times) and a distant third, the COCA (nine times).[31] The English UKWAC web corpus was used only four times, and the French FRWAC and the German DEWAC were each used once. Given the advantages of web corpora as argued for above, this is a baffling result. Before turning to an in-depth view of statistical methods used in my research, I therefore wish to point out that web corpora clearly seem to be underused in contemporary corpus linguistics. I hope the case studies presented below demonstrate their usefulness and inspire other corpus linguists to use them in their research.

2.3 Statistics

2.3.1 Overview

This section is exclusively about statistical modelling, which is often seen as an indispensable part of probabilistic modelling (Gries 2017a). Subsection 2.3.2 briefly describes known problems with inferential statistics as used by many practitioners. Subsection 2.3.3 discusses how certain prob-

[29] Additionally, a large-scale analysis of the distribution of automatically extracted lexicogrammatical features in web documents and the DeReKo corpus is being prepared for publication by the Institut für Deutsche Sprache and the present author.

[30] This is not the place to discuss technical details of web corpus construction. Schäfer & Bildhauer (2013) provides a convenient introduction to the subject.

[31] I do not even begin to discuss the problems of reproducibility involved when a custom corpus is created ad-hoc for a single research paper. The problem is even graver when web data is used in an ad-hoc fashion for corpus creation or even gathered by googling (Kilgarriff 2006).

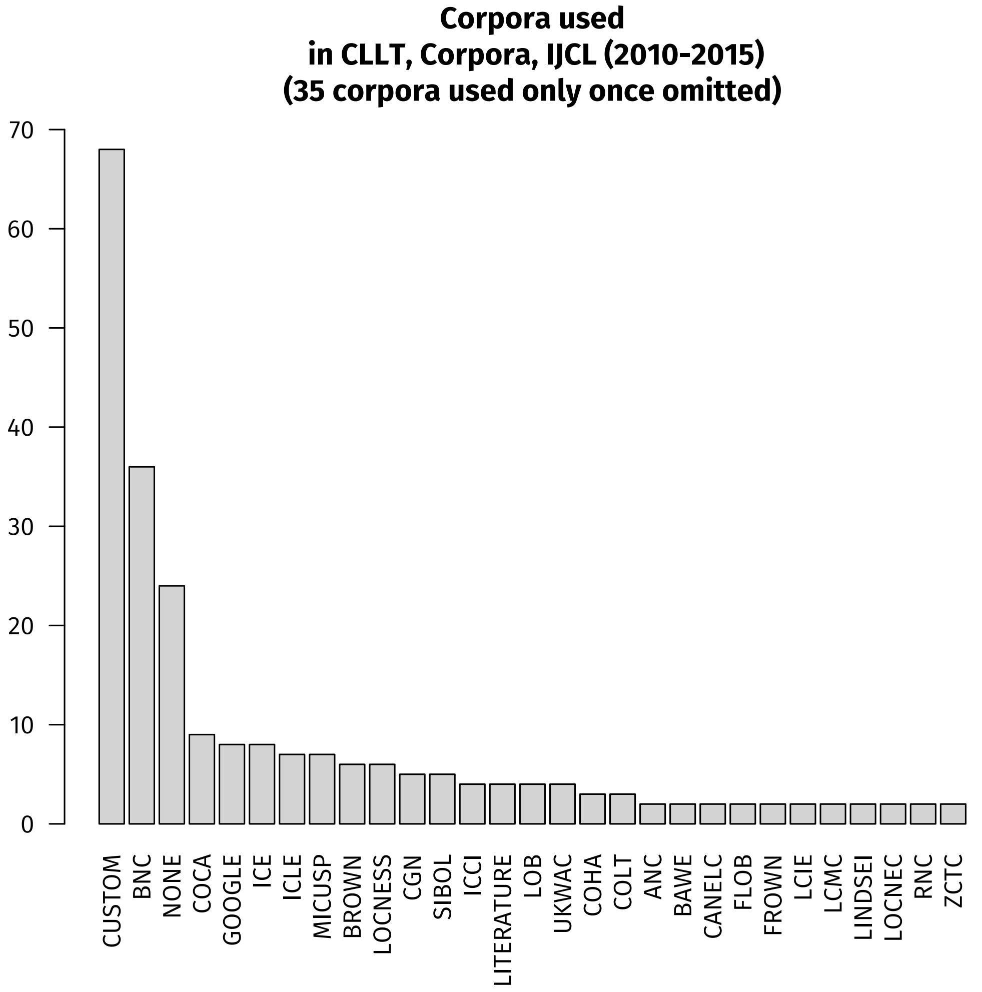

Figure 2: Corpora used in the three major corpus linguistics journals; CUSTOM was assigned when authors reported the creation of a custom corpus for the concrete study (a selection of newspaper articles, of academic papers, etc.); LITERATURE was assigned to papers about specific literary works; 111 corpora which were only used once (including DeReKo) are not shown.

lems with what is called the *non-randomness of linguistic data* do not affect many types of statistical analysis. Directly related to the previous subsection, Subsection 2.3.4 addresses problems with the associated idea that statistical models must always be exhaustive. Finally, Subsection 2.3.5 briefly shows that Bayesian modelling (as sometimes advocated for these days) usually will not lead to different results.

2.3.2 On statistical inference

In this section, I briefly discuss my position on data analysis and so-called *hypothesis testing*. The most widely used statistical system is *Null Hypothesis Significance Testing* (NHST), and it is one of the *frequentist* systems of statistical inference. In NHST, researchers attempt to substantiate the existence of an effect (such as a positive connection between the three different non-nominative cases on an NP and the occurrence of a non-standard form in the NP, see Schäfer 2016c) which is predicted to exist by their favoured theory by means of conducting an experiment in which the effect is measured. Then, the probability p (the so-called *p-value*) of obtaining the observed measurements or more extreme measurements under the assumption that there is actually *no* effect (the *null hypothesis* or just the *null*) is calculated. If this probability is lower than a certain threshold (usually called the *α-level*), the null hypothesis is *rejected*, which is taken as evidence that the hypothesis derived from the theory is correct. It is often incorrectly stated that *the experiment/test shows that the probability that the null is correct is* p or *is lower than* α. This approach is riddled with philosophical and statistical problems and has led to the promotion of of bad scientific practice. Among the most ardent critics are Gigerenzer (2004), Colquhoun (2014), and Munafò et al. (2017). The editors of the journal *Basic and Applied Social Psychology* have even banned the use of p-values in an actionist attempt to tackle problems of bad science related to NHST (Trafimow & Marks 2015). Critics often propose to abandon frequentist inference altogether and adopt a Bayesian approach, which itself is not without philosophical and practical problems (see, for example, Mayo 1996, Senn 2011). Others have proposed abandoning statistical inference proper in favour of confidence intervals and effect sizes (Cumming 2014), sometimes not noticing that NHST confidence intervals are not considerably different from NHST p-values, as Perezgonzalez (2015a) shows in reply to Cumming (2014).

However, there is no need to abandon frequentist inference or p-values simply because they have been abused. A great many statisticians and researchers have shown that the major problem with NHST is that it is a mixture of the statistical philosophies of Ronald A. Fisher on the one hand and Jerzy Neyman and Egon Pearson on the other hand (see Goodman 2008, Perezgonzalez 2014, Perezgonzalez 2015b, Greenland et al. 2016; see also Lehmann 1993 and Lehmann 2011 for an overview of these two philosophies and the history of their development). I follow Fisher's statistical philosophy, and I briefly compare it to Neyman and Pearson's now.

Neyman and Pearson developed a system where two hypotheses are specified: the *main hypothesis* ($\mathrm{H_M}$) and the *alternative hypothesis* ($\mathrm{H_A}$). These hypotheses have to exhaust the probability space such that $p(H_M \cup H_A) = 1$. The goal is to accept either of these hypotheses and reject the other, where typically $\mathrm{H_M}$ is the hypothesis predicted by the experimenter's favoured theory and the one they would like to accept. The reason why the Neyman-Pearson approach can be hard to implement is that $\mathrm{H_M}$ needs to be specified *precisely*, i. e.including the effect size. For example, if the experiment is a reading time experiment contrasting reading times under two distinct conditions, then the expected increase in reading times needs to be specified numerically. If this is possible, researchers can calculate the risk of incorrectly accepting $\mathrm{H_M}$ when it is false (α) and the risk of incorrectly accepting $\mathrm{H_A}$ when it is false (β) *given specific sample sizes*. Then, researchers can decide upon the optimal sample size and choose the optimal testing procedure. Especially Neyman designed this system explicitly with the idea in mind that researchers end up doing the right thing in $1-\alpha$ of all cases if they follow this protocol. No inference with respect to the ultimate truth of a specific hypothesis at hand was ever intended by Neyman, and all he wanted to achieve was long-run control of error rates.[32] I refer to this quote from Neyman & Pearson (1933: 290-291) on hypothesis testing:

> We are inclined to think that as far as a particular hypothesis is concerned, no test based upon a theory of probability can by itself provide any valuable evidence of the truth or falsehood of a hypothesis.

Similarly, Neyman (1937: 349) has this to say about frequentist confidence intervals (italics in the original):

> It will be noticed that in the above description the probability statements refer to the problems of estimation with which the

[32] Mayo (1996); Mayo & Spanos (2006); Mayo (2009; 2018) propose a theory of statistical inference (called *severe testing*) which is similar to the Neyman-Pearson system, but which also allows inferences about the case at hand. Unfortunately, severe testing is mostly uncharted territory for practitioners in most fields, including linguistics.

> statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results *will* tend to α. [fn. omitted, RS] Consider now the case when a sample, E', is already drawn and the calculations have given, say, $\underline{\theta}(E') = 1$ and $\bar{\theta}(E') = 2$. Can we say that in this particular case the probability of the true value of θ_1 falling between 1 and 2 is equal to α? The answer is obviously in the negative. The parameter θ_1 is an unknown constant and no probability statement concerning its value may be made […] .

In empirical linguistics (both corpus-based and experimental), following the Neyman-Pearson protocol is often impossible because theories do not predict effect sizes and/or no previous knowledge exists about the expected effect size.

Fisher developed a different system, in which the probability of a specific outcome (or a more extreme outcome) of a random experiment *if there is no effect* (the H_0 or *null hypothesis* or simply the *null*) is calculated as the p-value. It cannot be stressed enough that this is the probability of obtaining such results *before the experiment is conducted*, and that it is *conditioned on the design of the experiment*. It is *not* a Bayesian posterior probability which allegedly quantifies the credibility of a hypothesis given the data. Changing the experiment design changes the sample space and thus leads to different frequentist probabilities, even if the actual measurements are the same. Therefore, unrealised events play a role in the frequentist interpretation of experiments.[33]

Now, Fisher (1926: 504) suggests an informal, adaptive, and approximate *threshold of significance* (or *sig*), for example 0.05, below which researchers might suspect that there is something going on. While Fisher did not recommend the direct inspection or interpretation of p-values (at least not until very late in his life; see Section 4.4 of Lehmann 2011), he recommended that experimenters set *sig* appropriately based on previous experimental or theoretical knowledge. The most important pitfalls and misunderstandings (directly translating into some of the false assumptions common in NHST) in Fisher's framework are:

i. Researchers think that the p-value corresponds to the posterior probability (called the *inverse probability*; see Senn 2011) that the null

[33] This is seen as a problem by some statisticians and researchers who favour the likelihood principle (and Bayesian inference) over frequentism (Birnbaum 1962). See Mayo (2014) for a summary of the defense of frequentism against likelihoodism.

hypothesis is true. Or, even worse, they believe that the posterior probability that the substantive hypothesis is true is $1 - p$.

ii. Researchers take a significant result as a *proof* of something, usually the hypothesised effect. In fact, significance only shows that either the null does not describe the actual world very well *or a rare event has occurred*. There is no way of knowing with any specifiable accuracy which of these is the case.

iii. Practitioners take point (ii) even further and make an inference from a single significant result to some substantive hypothesis such as *my whole theory is correct*, forgetting that the test evaluates not just the theory, but also the adequacy of the experimental setup, the accuracy of the measurements, the operationalisations used to measure a theoretical construct, etc.

iv. Researchers assign high importance to some significant result and low importance to post-hoc effect size. This leads to overly optimistic interpretations of the data when they suggest that the null might be rejected ignoring that the effect is actually rather small.

v. If one runs a series of experiments and performs the corresponding tests in which the nulls are conceptually related, the actual probabilities of a rare event happening increase, and each p or the *sig* level are too optimistic if left uncorrected.

Point (i) has been addressed ad nauseam by statisticians and statistics-aware practitioners (see Goodman 2008, Perezgonzalez 2014, Perezgonzalez 2015b, Greenland et al. 2016). It is simply not true that frequentist p-values contain any information about the probability that any hypothesis is true given the evidence. The p-value (in Fisher's system, where p-values have a proper definition) is the probability of the outcome of the experiment (or a more extreme outcome) under the null *before the experiment was conducted*. After the experiment has been conducted, the outcome (however unlikely it might have been before the dice were rolled) is obviously factual and therefore has a probability of 1 like all other facts.

Points (ii) and (iii) can be remedied by researchers being aware of the (relatively) low importance which should be attributed to a single significant result. Furthermore, good use of previous experimental and theoretical knowledge in evaluating the actual p-values (although Fisher himself was not much interested in interpreting them) helps to make the Fisher approach more sound in practice. It also helps to do replications and perform meta-analyses. Problems with point (iv) are easily avoided by looking at post-hoc effect sizes. Fisher used the informal notion of *sensitivity* to alert practitioners that if, for example, a weak effect is detected with a very large sample, the result might not mean very much despite a successful rejection of the null. Demanding that researchers pay more attention to effect sizes is really just another way of saying that they should do proper exploratory/descriptive analysis of their data sets. Point (v) can be dealt with by applying corrections for group-wise error (which should not be called *α-level correction* under Fisher's approach, even if the two are mathematically equivalent).

One objection against Fisher-type statistical inferences comes from the underlying randomness assumption (see the second chapter of Maxwell & Delaney 2004 for a very accessible introduction to Fisher's ideas about randomness). Fisherian statistical inference is only valid if the randomness assumption (RA; see above in Section 2.2.1) holds. If practitioners do not conduct a proper random experiment (wilfully or out of ignorance), they are changing the sample space and thereby invalidating the actual computations of the statistical tests. This was addressed in a prominent paper on statistics for corpus linguistics, and the next section discusses this problem.

2.3.3 Language is never ever random?

In Kilgarriff (2005), Adam Kilgarriff made an interesting argument about corpora, statistics, and the RA. In this section, I briefly review the most important points of his argument and propose (in a point-by-point fashion) that the problems mentioned by Kilgarriff do not affect the type of research presented here more than any empirical science.

First of all, Kilgarriff points out that the relation between two types of events (such as the occurrences of two lexemes next to each other in a corpus) can be one of the following (with my paraphrases of the terms' interpretations).

- **random**
  completely uncorrelated
- **arbitrary**
  co-occurring without an underlying causal relationship
- **motivated**
  co-occurring because of an underlying causal relationship
- **predictable**
  standing in a causal relationship where one event is a sufficient condition for the other

The strength of the link between events standing in these types of relationships to one another obviously increases from top to bottom. Kilgarriff explains correctly that significance testing is only able to discern between a situation of randomness (R, the situation under the assumption of the null hypothesis) and non-randomness (¬R).[34] Whenever two types of events do not stand in a random relationship to one another (such as weak nouns occuring more often in a strong form when the NP has dative case compared to when it has genitive case, see Schäfer 2016c), no statistical system (including Bayesian statistics) can help us to decide whether the correlation is arbitrary (or *accidental*, to use a less precise colloquiual term) or motivated, i. e. causal. While Kilgarriff is absolutely right in pointing this out, the situation is exactly the same in any science. This is why hypotheses are usually chosen with great care and based on sophisticated theories (see Chalmers 2013 for an introduction, especially chapters 5–7). It is essential to test only substantive hypotheses and give the test the best and toughest chances of finding errors in the theory which generated the hypotheses. In other words, while Kilgarriff is entirely right, it is also completely unjustified to expect statistics to do the job that theory and experiment design usually do. To illustrate this point further, Kilgarriff's example goes like this: A study might find that shoe polish and cat food are bought simultaneously significantly more often than to be expected under the null (i. e. given the number of times the items are bought). A statistical test might reject the null, which states that *the probability of shoe polish being bought when cat food is also bought is the same as the probability that shoe polish is bought when no cat food is bought*. However, we would have no reason to assume that the relationship between the two types of events would be anything but arbitrary, although it would be very likely not random. He argues that this could be accounted for if both articles are more often bought together for independent reasons during hot weather or something along those lines. The point is that any researcher who would conduct such an experiment *without a substantive theory-driven hypothesis* about why buying shoe polish and buying cat food should be correlated has already engaged in bad science, and frequentist statistics cannot be blamed for this.

A second major point raised in Kilgarriff (2005: 266) is that

> [w]hether we can reject the null hypothesis […] is a function of the sample size and the level of correlation. Where sample

[34] Kilgarriff consistently uses the term *NHST* without making it clear whether he means NHST in the narrow sense described above or Fisherian inference. It appears clear to me that he does not have Neyman-Pearson error control in mind.

> size is held constant (and is not enormous), whether or not we can reject H_0 can be seen as a way of providing statistical support for distinguishing the arbitrary and the motivated. This is a role that hypothesis testing plays across the social sciences.

This, too, is undoubtedly true. It is a known triviality that whenever the sample size is large enough, any minor (and arbitrary in the sense explained above) correlation leads to a significant test result. Kilgarriff argues that language is never random, in the sense that grammar, semantics, pragmatics, etc. always cause words to be chained together in a non-random fashion, and that it should be expected that in large corpora virtually any co-occurrence of words will turn out to significantly contradict the null. At the same time, these significant co-occurrences of words will often be arbitrary, i. e. merely an accidental result of theoretically meaningless interactions of the non-random mechanisms of grammar. While this is also very true, it is neither specific to linguistics nor is it an argument against significance testing per se. It is an argument against the search for significant results unguided by concrete theoretical knowledge and without paying attention to setting the appropriate *sig* levels or to the sensitivity of the test (all in Fisher's terms, see Section 2.3.2). If researchers in social sciences simply searched large databases containing socioeconomic data for correlations between variables, numerous correlations would be detected as significant which at the same time would be completely arbitrary and even entertainingly funny.[35]

At this point, we have to consider where Adam Kilgarriff is coming from. His whole argument indicates that he is thinking in terms of collocation research, and the statistical measures he discusses later in the paper confirm this assumption. His paper is, however, often and prominently referenced in a global fashion when corpora and problems of statistical inference are discussed (not just with reference to research on collocation), for example in Divjak, Dąbrowska & Arppe (2016: 2). I argue that there are substantial differences between research on *collo* phenomena and alternation modelling. In collocation research, it is customary to examine huge numbers of pairs of words to find those which co-occur disproportionally often with each other in a certain window of, for example, five words. Despite the fact that this is a form of data analysis and not a type of theory-driven testing of single substantive hypotheses in well-crafted experiments, measures of evidence (i. e. hypothesis testing) are sometimes used to find *significant collocates* (see Ev-

[35] The book Vigen (2015) is an amusing illustration of such spurious correlations.

ert 2008 for an overview including criticism of such measures).[36] Kilgarriff is perfectly right in pointing out all of these shortcomings, but his criticism simply does not apply to the type of work presented here.

For corpus studies like the ones presented here, the unrestricted query approach described in Section 2.2 leads to samples containing the relevant constructions or alternants as they appear in the corpus. The samples are then annotated (often manually) for a number of theoretically well-founded *regressors* (independent variables, also called *predictors*) and a *response variable* (dependent variable, also called the *outcome*), which is (in the case of a binary alternation) simply a binary variable encoding the choice of the alternant. This is substantially different from collocation research, where arbitrary sequences of words are examined for unusually high co-occurrence frequencies.[37] The default assumption (the null) in alternation research is indeed that any regressor (all other things being random) does not correlate with the response variable, i. e. the choice of the alternant. In other words, the arbitrary (but non-random) influence of grammar, which mars hypothesis testing for collocations as pointed out by Kilgarriff, is eliminated, because the study focusses on a very narrowly defined grammatical configuration anyway. Furthermore, samples are usually of a moderate size (several thousands of exemplars) given the complexity of the multifactorial statistical models, such that the tests have a reasonable level of sensitivity.

2.3.4 Model everything?

This section discusses some technical points related to the statistical models used in contemporary alternation modelling. Readers with a background in statistics or those not interested in in-depth statistical discussions are invited to skip it.

[36] The same was tried in collostructional analysis (see Stefanowitsch & Gries 2003; Gries & Stefanowitsch 2004). While this was still criticised in Schmid & Küchenhoff (2013); Küchenhoff & Schmid (2015), collostructional approaches have actually been moving away from measures of evidence to measures of effect strength Gries (2012; 2015a).

[37] Using an analogy from the social sciences, the difference would resemble that between, on the one hand, looking for correlations between arbitrary socioeconomic factors in census data, and on the other hand, a specific correlation between peoples' voting behaviour in general elections and certain socioeconomic factors which are known to influence voting behaviour. Both types of data could be drawn from the same large existing database, but the studies differ significantly in their theoretical well-foundedness and the sampling scheme.

Introduction In my research, generalised (mixed/hierarchical) linear models (i. e. some form of regression) are used as a de facto standard. Looking at the frequency with which statistical procedures are applied in the three major corpus linguistics journals (according to the survey to be published as part of Schäfer n.d.), regression is the most prominent advanced multifactorial statistical method used in corpus linguistics. See Figure 3, which shows that simple descriptive statistics (appearing 84 times) and monofactorial methods like the likelihood ratio test (LLR; 32 times) and the χ^2 test (CHISQ; 31 times) are still dominant, but that regression comes in fifth with 22 uses (198 papers in total with 378 distinct uses of statistical methods).

With regression models becoming (at least a part of) the state of the art in corpus linguistics, I want to point out that a recent trend to *model everything* might be justified but not strictly required. First, I provide a brief introduction to regression modelling using binary regression, a highly popular type of regression in the modelling of binary alternations, as an example.[38] Then, I discuss the programme laid out in Gries (2017b), where a number of factors are enumerated that regression models *should* take into account.

Generalised linear models Binary regression (usually logistic regression) models the influence a number of independent variables cumulatively exert on a binary dependent variable.[39] In the regression literature, the independent variables are usually called *regressors* and the dependent variable is called the *response*. In Schäfer (2018), for example, the dependent variable was 0 when the alternant given here as (1-a), where the kind-denoting noun (*Wein*) and the measure noun (*Glas*) agree in case, was chosen in the exemplar, and it was 1 when the alternant given here as (1-b), where the kind-denoting noun has genitive case, was chosen.

(1) a. Wir trinken [[ein Glas]$_{\text{Acc}}$ [guten Wein]$_{\text{Acc}}$]$_{\text{Acc}}$.
we drink a glass good wine
We drink a glass of good wine.
b. Wir trinken [[ein Glas]$_{\text{Acc}}$ [guten Weins]$_{\text{Gen}}$]$_{\text{Acc}}$.

The regression estimates the influence of any number of predictors on the probability that the response is 0 or 1. The theoretically motivated regressors in this particular study included the case of the measure head noun, the

[38] My view on regression and multilevel models is strongly guided by Gelman & Hill (2006). I also use Zuur et al. (2009); Fahrmeir et al. (2013); Fox (2016) as reference text books.

[39] This short introduction is partially based on Schäfer (2019).

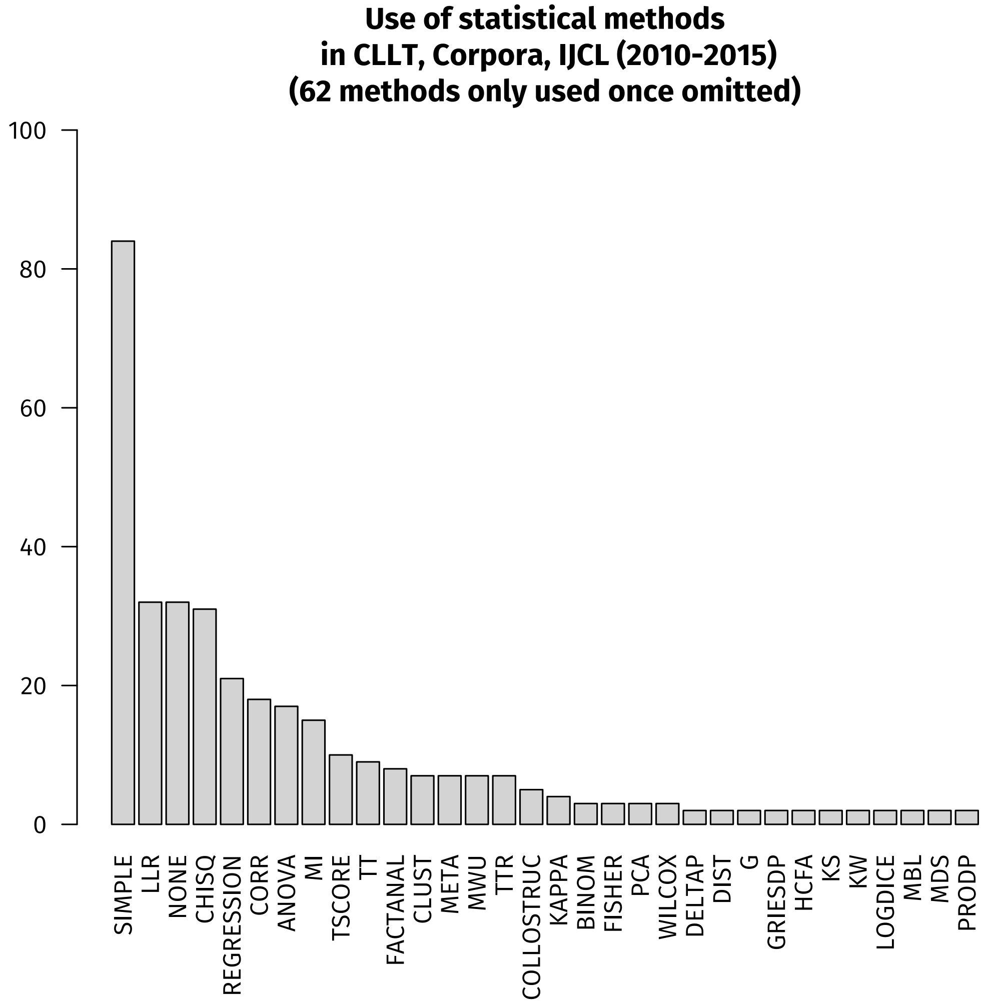

Figure 3: Use of statistical methods in major corpus linguistics journals; multiple codes per paper were possible; SIMPLE was coded for papers using descriptive statistics, plots, or simple comparisons of relative or absolute frequencies; NONE was coded for papers using no statistics at all; META was coded for methodological papers (in which case individual methods were *not* coded).

semantic class of the measure noun, lemma frequencies of both the measure noun and the kind-denoting noun, document-level indicators of style, and the type of determiner (cardinal or not) used for the whole NP. The regressors are thus binary, nominal, and numeric.

Generalised Linear Models all work by assuming that the concrete measurements of the features assumed to influence the outcome are multiplied by *coefficients* which encode the direction (positive or negative) and the strength of each regressor's influence. Continuing with the above example, if the NP in an exemplar indexed i contains a cardinal determiner, the measurement would be $x_i^c = 1$, and the model specification would be such that this term is multiplied by a coefficient β^c to yield a numeric quantification of the influence (all other things being equal) on the decision to use either variant. Thus, the model sub-term for this particular regressor looks like (2).

(2)
$$x_i^c \cdot \beta^c$$

The sub-terms for all m regressors look the same, and they are added up to form the *linear term*. Also, the *intercept* α^0 is added, which encodes the resulting value when all other regressors are 0 and consequently lead to their corresponding sub-terms being 0. (3) shows a general form for the linear term with m sub-terms. The index i indexes the single observations (given sample size n, $i \in \{1..n\}$).

(3)
$$\alpha^0 + \beta^1 \cdot x_i^1 + \cdots + \beta^m \cdot x_i^m$$

The influences on the response variable are thus added up, and they are all assumed to encode a linear relationship. However, such a linear term can assume arbitrary positive and negative values. A probability, which is what we want to model, is always in the interval $[0,1]$. To turn the result of the linear term into a probability, the *link function* is applied to the linear term. For binary regression, the inverse logit or probit functions are typically used, such that the full model specification looks like the form in (4).

(4)
$$Pr(y=1) = logit^{-1}\left[\alpha^0 + \beta^1 \cdot x_i^1 + \cdots + \beta^m \cdot x_i^m\right]$$

(4) specifies a model and encodes that the Probability Pr of the response y being 1 – or simply $Pr(y = 1)$ in mathematical notation – is given by

the inverse logit $logit^1[]$ of the linear term which consists of the overall intercept α^0 plus the added up concrete values of each x_i^j multiplied by the corresponding coefficient β^j, in other words $\beta^j \cdot x_i^j$. When we specify such a model, we ideally make a theoretical commitment to the factors that drive the choice of the alternants. Setting up the model is thus the crucial step in going from theoretical considerations to a quantitative analysis.

The job of the so-called *estimator* (a type of algorithm typically implemented in statistics software) is to find the optimal values of all β^j given the observations (the annotated sample). Given these values of the coefficients, the model will predict with the best possible accuracy the probabilities of alternant choices given any assignment of the regressor values. Ideally, the coefficients would be estimated in a way such that the model predicts each outcome encountered in the sample perfectly.[40] This is virtually never the case, and there is always going to be a difference between the predictions and the actual outcomes. These errors are presumed to follow a specific distribution, which is an assumption underlying the estimation process. In the case of binary regression, the distribution of errors is assumed to be the binomial, and the model presented here would therefore be called a *binomial generalised linear model with a logit link function*. Generalised linear models are abbreviated as GLM.[41]

Random effects An extension of GLMs are *generalised linear mixed models* (GLMMs) or simply *multilevel GLMs*. The difference is that GLMMs contain so-called *random effects*. To understand a random effect, a *random intercept* is the best point of departure. Any nominal variable like grammatical case or verb lemma or speaker has a certain number of levels. In the following illustration, l is used to denote this number. Each level defines a group of exemplars (such as those in the nominative, those with the verb *give*, or those uttered by a specific speaker), and they can therefore be called *grouping factors*. What happens if we use such a variable as a fixed-effect predictor in a GLM (instead of a random effect in a GLMM) is *dummy coding*.

Dummy coding (or *contrast coding*) is a way of encoding a categorical variable as a number of binary variables. See Table 1 for an illustration of how German case (a four-way variable) could be dummy coded. The l levels of the grouping factor are dummy-coded as $l-1$ binary variables.

[40] I. e. a probability of 1 would be predicted when the actual outcome was 1, and a probability of 0 would be predicted when the actual outcome was 0.

[41] Actually, there is a technical distinction to be made between the logistic regression introduced here, which models probabilities, and a proper binomial GLM, which models counts using a binomial regression. The difference can be neglected for the present purpose.

Actual variable	Dummy variables		
Case	**accusative**	**dative**	**genitive**
Nominative	0	0	0
Accusative	1	0	0
Dative	0	1	0
Genitive	0	0	1

Table 1: Dummy coding of a categorical variable *Case* with four levels, resulting in the three binary dummy variables *accusative*, *dative*, *genitive*.

Since the first of the l levels of the grouping factor is encoded by all dummy variables assuming the value 0, only $l-1$ sub-terms are added to the model, and consequently only $l-1$ coefficients are estimated. The first level of the actual nominal variable (*Nominative* in the example) is thus *on the intercept* and becomes the reference to which all other levels are compared.[42] In such a model, the effect of each grammatical case is treated as a fixed population parameter, and one coefficient is estimated for each dummy case. In other words, the algorithm which estimates the coefficients for the $l-1$ dummy variables tries to find a fixed value for each of them without taking the variation between them into account. With many levels, this requires a lot of data, and levels for which only a few observations are available in the data set have very imprecise coefficient estimates with large confidence intervals.

Random intercepts are a way of using grouping factors without dummy coding and by taking the between-group variance into account. They are not estimates of fixed population parameters (*fixed effects*) but predictions of random variables. If we treat a grouping factor as a random intercept, we simply let the intercept vary by group (by adding a group-specific constant to the overall intercept), and we give the varying intercepts a distribution instead of estimating $l-1$ coefficients. This is the relevant difference between a fixed effect and a random effect. The general model specification with one random intercept looks like (5).

$$Pr(y=1) = logit^{-1}\left[\alpha^0 + \alpha^1_{g[i]} + \beta^1 \cdot x^1_i + \cdots + \beta^m \cdot x^m_i\right] \tag{5}$$

[42] Picking one dummy as a reference level is necessary because otherwise, infinitely many equivalent estimates of the model coefficients exist, as one could simply add any arbitrary constant to the intercept and shift the other coefficients accordingly. However, the estimator works under the assumption that there is a unique maximum likelihood estimate.

The only addition compared to (4) is $\alpha^1_{g[i]}$. I use the notation $g[i]$ (borrowed in a modified form from Gelman & Hill 2006) to indicate that the appropriate g-th lemma intercept is chosen for the i-th observation. If, for example, exemplar 9 contains the verb *give*, which is encoded as group 13, then $i = 9$, and $g[i] = g[9] = 13$. Thus, we now have an intercept which varies by group (instead of one term with its own coefficient per group). Crucially, instead of estimating a batch of coefficients for the lemma effect, the random effect is itself modeled, and random terms are predicted for each level of the random effect. For this, the assumption in (6) is made.

(6) $$\alpha_g \sim N(\mu_g, \sigma^2_g)$$

This is standard notation to indicate that the values of α_g follow a normal distribution with mean μ_g and a variance of σ^2_g. In fact, we can regard (6) as a minimal second-level linear model already, although one which simply predicts varying intercepts from a normal distribution. All more complex mixed or multilevel models are extensions of this approach.

Choosing fixed or random effects The decision between a fixed (dummy-coded) effect and a varying intercept boils down to two points. First, the variance in the intercepts needs to be estimated. Second, the random intercepts can be understood as a compromise between fitting separate models for each group of the grouping factor (*no pooling*) and fitting a model while ignoring the grouping factor altogether (*complete pooling*); see Gelman & Hill (2006: Ch. 12). As was stated above in (6), the random intercepts are assumed to follow a normal distribution, and the variance σ^2_g needs to be estimated with sufficient precision. From the estimated variance and the data, the estimator then predicts the *conditional modes* in GLMMs for each group (see Bates 2010: Ch. 1), which is the numerical value for each group which encodes the per-group tendency. This procedure, however, requires that the number of groups must not be too low to effectively achieve this. As a rule of thumb, having fewer than five levels means that a grouping factor should be included as a fixed effect, regardless of its conceptual nature. Even if there is a default recommendation to use a speaker grouping variable as a random effect, it is ill-advised to do so if there are exemplars from less than five speakers in the sample. Along the same lines, mode (typically spoken vs. written) is no suitable grouping factor for use as a random effect. Very often, the estimator will simply fail under such conditions, and a fixed effect might be the only option for technical reasons.

If, however, the number of groups is reasonably large, the number of observations per group is the second parameter to consider. Alternatives to using a random effect would be to estimate a separate model for each level of the grouping factor or to include it as a fixed effect. If a random effect is used, the conditional modes are *shrunken* (i. e. pulled) towards the overall intercept. This is called *shrinkage*. When the number of observations in a group is low, the conditional mode is shrunken more strongly, and only a small deviation from the overall tendency is predicted for the group. In such a situation (low numbers of observations per group), fixed effect estimates would turn out to be inexact. Clearly, low numbers of observations in all or some groups speak against using fixed effects grouping factors. Random effects are unproblematic under such conditions because of shrinkage.

This purely technical view of fixed vs. random effects is not in line with most introductory textbooks written for practitioners. One commonly given reason for using a random effect instead of a fixed effect is that the researcher is allegedly not interested in the individual levels of the random effect or similar, seemingly conceptual arguments. It appears that there is little foundation to such claims. Gelman & Hill (2006: 245–247) summarise the diverging and contradictory recommendations regarding what should be a random effect as found in the literature. They conclude that there is essentially no universally accepted and tenable conceptual criterion for deciding what should be a random effect and what a fixed effect. I agree with them and consider the decision primarily a technical one, i. e. we use what works best, preferring random effects whenever possible.

Varying intercepts and slopes If the coefficients β also vary by group, varying slopes are a possible extension of simple GLMs. We extend the model from (5) by giving β^1 a random slope. Instead of estimating a fixed coefficient, coefficients are predicted and assumed to come from a random (normal) distribution. The other fixed effect coefficients remain the same; see (7). We use $\beta^j_{g[i]}$ to denote the coefficient j varying by group g, which is chosen appropriately for exemplar i.

(7)

$$Pr(y=1) = logit^{-1}\left[\alpha^0 + \alpha^1_{g[i]} + (\beta^1 + \beta^1_{g[i]}) \cdot x^1_i + \beta^2 \cdot x^2_i + \cdots + \beta^m \cdot x^m_i\right]$$

A source of problems in varying intercepts and varying slopes (VIVS) models is the fact that in addition to the variance in the intercepts and slopes, the covariance between them has to be estimated. If in groups with a higher-than-average intercept, the slope is also higher than average, then they are positively correlated, and the reverse applies for lower-than-average intercepts and slopes. These relations are captured in the *covariance*. Condition (8) is added (the superscript indices on α and β have been omitted for readability).

$$\begin{pmatrix}\alpha\\ \beta\end{pmatrix} \sim \left(\begin{pmatrix}\mu_\alpha\\ \mu_\beta\end{pmatrix}, \begin{pmatrix}\sigma_\alpha^2 & \rho\sigma_\alpha\sigma_\beta\\ \rho\sigma_\alpha\sigma_\beta & \sigma_\beta^2\end{pmatrix}\right) \tag{8}$$

(8) says that the joint distribution of the intercepts α and the slopes β follows a bivariate normal distribution with means μ_α and μ_β. The variance in the intercepts is σ_α, the variance in the slopes is σ_β, and the coefficient for the covariance between them is ρ. Figure 4 shows the bivariate density distributions for two (1) negatively correlated, (2) non-correlated, and (3) positively correlated normally distributed variables.

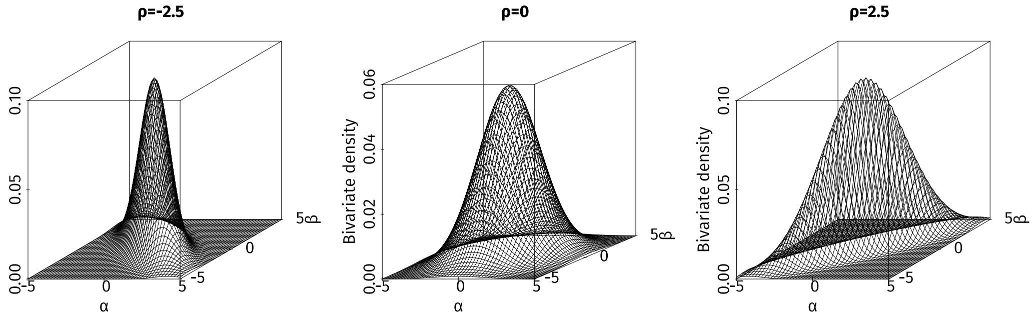

Figure 4: Bivariate normal density distribution with different correlation coefficients ρ; $\sigma_\alpha = \sigma_\beta = 3$; $\mu_\alpha = \mu_\beta = 0$.

The number of variance parameters to be estimated thus obviously increases with more complex model specifications, and the estimation of the parameters in the presence of complex variance-covariance matrices requires considerably more data than estimating a single variance parameter. The estimator might converge, but typically covariance estimates of -1 or 1 indicate that the data was too sparse for a successful estimation of the parameter. In this case, the model is *over-parametrised* and needs to be simplified. See Bates et al. (2015); Matuschek et al. (2017).

Second-level predictors The linear Gaussian models for random intercepts and slopes can also have fixed-effect regressors themselves (see Schäfer 2018 for an application). This means that the random effects are partially predicted from a set of separate fixed-effect regressors. A good example are per-lemma random effects to take care of lemma-specific preferences, but with lemma frequencies, semantic classes of the lemmas, etc. being able to partially predict these tendencies. Thus, the tendencies are not just idiosyncrasies of lemmas, but also determined by properties of the lemma. In this case, an additional linear model is specified for the random effect instead of the simple normal distribution predictor. We now extend (5) by a predictor δ_1 as a second-level predictor for α^1_g. The first-level model specification remains the same, and it is repeated here as (9).

(9)
$$Pr(y=1) = logit^{-1}\left[\alpha^0 + \alpha^1_{g[i]} + \beta^1 \cdot x^1_i + \cdots + \beta^m \cdot x^m_i\right]$$

However, instead of (6), the varying intercept is now predicted from (10).

(10)
$$\alpha^1_g \sim N(\gamma^0 + \delta^1 \cdot u^1_g, \sigma^2_g)$$

Instead of just predicting the mode of each α^1_g value, the model in (10) specifies a second-level intercept γ^0 and a second-level fixed coefficient δ^1, where u^1_g is the value of the second-level regressor variable for group g. True multilevel models increase the complexity of GLMMs, especially if third-, fourth-, or more-level models are used. Situations for multilevel modeling are quite frequently encountered. Especially when it comes to speakers as random effects, age, gender, region of birth (if this grouping factor has too few levels to be used as a random effect nesting the speaker random effect), etc. are ideal second-level predictors. The same goes for lemma frequencies, semantic classes, etc. as pointed out above.

This short introduction demonstrates that GLMMs or multilevel models can quickly become highly complex. With increasing complexity, however, more and more data are required for the estimation of the parameters and the predictions of the random variables. This leads nicely into the final point I want to make in this section.

Modelling everything So far, this section has provided a minimal introduction to GLMMs and true multilevel models. I have shown that models become rather complex relatively easily, and it was pointed out that especially in VIVS models, the necessary estimation of variance-covariance matrices requires a lot of data. Over-parametrisation in models for experimental data (as touted by Barr et al. 2013) was heavily criticised by Bates et al. (2015); Matuschek et al. (2017). In a situation of over-parametrisation, it should be noted that even an estimator which is more robust (such as maybe Bayesian estimators) cannot make reliable inferences possible where the data are insufficient given the model's complexity.[43] Bates et al. (2015: 1) state (emphasis is mine):

> We show that failure to converge typically is not due to a suboptimal estimation algorithm, but is a consequence of attempting to fit a model that is too complex to be properly supported by the data, *irrespective of whether estimation is based on maximum likelihood or on Bayesian hierarchical modelling with uninformative or weakly informative priors*. Importantly, even under convergence, over-parametrisation may lead to uninterpretable models.

For corpus-based alternation research, Gries (2017b) argues for massively more complex models compared to the current state of the art (including his own work), and over-parametrisation will certainly occur in many cases. His line of argumentation follows the previously published Gries (2015b). In Gries (2017b: 21–24), he argues that a large number of predictors should be standardly added to alternation models, such as lengths of words and constituents, information-structural predictors, NP types, animacy, and priming/persistency predictors, which encode potentially numerous effects of words and constructions which occurred before the target item. In Gries (2015b), he argues that random effects for text types, speakers, lemmas, etc. should also be added as part of the standard protocol (at least whenever the corpus contains the appropriate metadata). Gries (2017b: 24) admits that

> [t]he bad news, as so often, is that this requires an ever increasing or nearly overwhelming degree of sophistication and knowledge not only in linguistics but also in matters of data analysis and statistical evaluation.

Given that complex models are used in many fields, I would argue that what can be expected of biologists, sociologists, econometricians, etc. (or students of these fields) can be expected of linguists, too. Textbooks at all levels of technicality exist, and Gelman & Hill (2006) is a true eye-opener which is still accessible for practitioners. My main practical objection to the

[43] Read more comments on Bayesian estimators in Section 2.3.5.

model everything approach (MEA) is another one. Given the complexity of the resulting multilevel models and the fact that a great many of the variables mentioned by Gries (2017b) would have to be annotated manually for any given study, the time and manpower required to conduct even a simple experiment would be disproportional, especially in a competitive academic environment where researchers are pressured to publish several studies each year. After all, the ideal models under Gries' MEA would very likely require samples containing many tens of thousands of exemplars.

It would be highly frustrating, however, not to implement MEA simply because it is not feasible due to limited resources. Therefore, I would like to argue that it is often not even necessary for substantive reasons. MEA is guided by very good reasons, and I will now discuss the two most important ones, interspersed with discussions of why, despite these reasons, this approach may still be unnecessary. First of all, the goal in alternation modelling might be to model the cognitive representations and processes which make speakers use one form or the other. In this case, we should indeed add *all relevant causal* predictors available (with emphasis on *all*, *relevant*, and *causal*). However, unless we have a tight and formally precise theoretical model of these cognitive representations and processes, simply adding everything that we can think of to the model which *might* influence the choice of alternants (even if other studies have produced evidence for such influences under different circumstances) is misguided and prone to producing a significant number of spurious correlations (see Section 2.3.3). This point is strongly related to the limited power corpus data have in making inferences about cognitive representations and processes (see Sections 2.1 and 2.2). In my studies, much like in controlled experiments, I use models which specify a configuration of theoretically motivated predictors, and I do not include any predictors (except sometimes lemma-specific random effects and lemma frequency) for which there is no substantive hypothesis or at least sound previous research.[44] Under MEA, it is even possible that an arbitrary but strong effect masks a substantive and causal effect (see Section 2.3.3 for a discussion of these terms), and the danger of this happening rises with the number of predictors added to the model on a hunch. Put differently and more radically, doing a monofactorial study which substantiates the existence of a causal mechanism between a predictor and the alternation might

[44] In Schäfer (2018), for example, previous independent theory-driven research by Zimmer (2015) had shown that grammatical case influences the alternation. Although my own theoretical model did not include case as a predictor, I still added it to increase the quality of the model fit (see below) without risking picking up spurious effects.

be worth substantially more than estimating a complex model (maybe even with high predictive power) that is not substantive.[45]

The second argument for MEA is that unmodelled variance in a (G)L(M)M has negative effects on the estimator. Some textbooks like Zuur et al. (2009) capitalise on this. Depending on the severity of this effect, models which are not maximal could actually be wrong, and this argument certainly deserves closer examination. Two types of situations have to be considered in this regard. In the first type of situation, an omitted predictor interacts with one or more predictors in the model. For example, the discourse status of the subject is part of the model specification, but the lemma of the governing verb is not. However, in reality, the two variables interact, meaning that the strength of the influence of the discourse status is different for different governing verbs. In this case, the actual coefficient estimates will be different for the predictors which are part of the model specification depending on whether the other regressor (verb lemma in the example) is included or not. There is no principled solution to this problem, however. First of all, simply adding all conceivable types of regressors (MEA) and hunting for interactions is, again, just a recipe for finding many spurious effects. Second, the coefficients might change if some predictor is added, but if its effect and the interaction with another predictor are arbitrary and not motivated (i. e. causal), then we should not be interested in the *updated* coefficients. Again, working with substantive models is the solution to this problem.

In the second type of situation, the omitted predictor and the non-omitted predictors do not interact. In this case, the coefficient estimates for the non-omitted predictors will be stable, i. e. unaffected. However, due to excess variance, the estimates of the standard deviations for the fixed-effects coefficients will be different, which is why Barr et al. (2013) predict dramatically inflated type I error rates for non-maximal models.[46] This effect is often used to argue for the inclusion of random effects (Gries 2015b) although it does not matter at all whether the omitted predictors are used as random or fixed effects (see above on the difference). However, as long as we do not over-parametrise our models, we have so much data that Fisherian tests on fixed effect coefficients are highly sensitive, and p-values for strong effects are mostly extremely low (see all four case studies collected here). Thus, adapting the standard *sig* level for alternation studies using

[45] Of course, things get significantly worse if a MEA model is taken as evidence for causal mechanisms. In no way, shape, or form does Gries do this in any of his publications, but other researchers might (implicitly or explicitly).

[46] A type I error is a term from the Neyman-Pearson theory of statistical inference, often used in NHST as well. It occurs when the null is true but rejected by the test.

GLMMs will take care of this problem. Substantial effects will usually still reach *sig* = 0.001, but unmodelled heterogeneity will not cause insubstantial effects to reach *sig* = 0.001. Conversely, with over-parametrised models, *sig* = 0.05, and even more data, we run once again the risk of detecting spurious effects and ending up with essentially uninterpretable models (Bates et al. 2015; Matuschek et al. 2017).

A minor nuisance in the second type of situation is that in GL(M)Ms with a binary response (logit and probit models) unmodelled variance (or *unmodelled heterogeneity*) pulls the coefficient estimates towards 0, which is called the *attenuation bias* (Wooldridge 2010: 582–585). However, our samples are usually large enough in corpus linguistics (at least – again and ironically – as long as we do not over-parametrise the models), and

> we should remember that, in nonlinear models, we usually want to estimate partial effects and not just parameters. For the purposes of obtaining the directions of the effects or the relative effects of the continuous explanatory variables, estimating β/σ [the coefficient biased towards 0 through unmodelled heterogeneity; RS] is just as good as estimating β [the true coefficient; RS].
>
> To be more precise, the scaled coefficient, β_j/σ, has the same sign as β_j, and so we will correctly (with enough data) determine the direction of the partial effect of any variable – discrete, continuous, or some mixture – by estimating the scaled coefficients. (Wooldridge 2010: 583)

Thus, it will be harder to detect an effect if there is a significant attenuation bias, but we will still make very similar inferences given enough data.

To summarise, I have argued for well-specified, non-maximal models and against MEA. The dangers of uninterpretable over-parametrised models which attempt to model many more effects than the amount of data is appropriate for by far outweigh the risks of omitted regressors or unmodelled variance. Furthermore, MEA might lead to many spurious effects being included in models, and thus probably being taken seriously. Instead, I argue for model specifications grounded in substantive theory and previous research. If such types of model specifications are difficult to come up with, I suggest the field invest more time and resources into producing powerful, highly predictive, and formally specified theories instead of engaging in data dredging.

2.3.5 Just go Bayesian?

In this section, I justify the choice of statistics which I used in all my studies collected here, namely non-Bayesian statistics. Over the past few years, modified versions of or alternatives to (multilevel) generalised linear models with a Maximum Likelihood Estimator (MLE) have been proposed. From among these methods, I just make a few remarks on Bayesian estimation (see Gelman, Carlin, et al. 2014) as it was proposed in Levshina (2016) and Divjak (2016), for example. Conceptually, I see three points of discussion that should be kept apart. First, Bayesian methods are sometimes touted as superior tools for scientific inference compared to frequentist methods. Second, it has been proposed that the Bayesian interpretation of probability is more cognitively adequate for the modeling of linguistic data (Divjak 2016: 301–302). Third and very specific to this paper, given the established methods in the modeling of alternation and variation, it has to be decided whether so-called Bayesian methods lead to substantially different results.

As for the first point, the relevant fundamentals of frequentism have already been mentioned in Section 2.3.2. The basic distinction between frequentism and Bayesianism is a philosophical one and related to the concepts of *direct* and *inverse probability* (e. g. Senn 2011). Frequentists assume that models and parameters are fixed and given by theories, for example a model specifying that a coin is fair. We can then calculate for observed data (for example a measurement of three heads in an experiment with ten tosses) how often such a result or a more extreme result would occur if the model were true and if we repeated the experiment arbitrarily often. This is essentially the frequentist notion of direct probability, i. e. long-run frequencies under replication. Standard tests in the Fisher and Neyman-Pearson traditions as well as Neyman confidence intervals are based on this concept of probability. Bayesian approaches (in the now common interpretation), on the other hand, are conditioned on the particular data and quantify inductively the probability of model parameters given the available data. The parameters are thus not fixed, and the resulting probability is usually equated with researchers' posterior beliefs about model parameters. The problem is that researchers often need a criterion that tells them whether a hypothesis was substantiated by an experiment or not (hypothesis testing and error control). There is actually a debate among Bayesians about the proper interpretation of Bayesian methods and whether a notion of hypothesis testing is compatible (or even already contained) in the Bayesian approach. In Gelman & Shalizi (2013: 10), the authors – prominent Bayesians themselves – acknowledge that a theory of statistical testing is a desideratum, state about the standard inductive interpretation of Bayesianism that "most of this received view of Bayesian inference is wrong", and develop a Bayesian notion of p-values (see also Mayo 2013, for a frequentist reply; also Senn 2011 on different strands of Bayesianism and their stance on inductive vs. deductive reasoning, and Mayo 2011, for a critical reply to Senn 2011). Clearly, in such quarrels between and among camps of science philosophers and statisticians, it is difficult for mere practitioners to take sides.

Turning to the second point, Divjak (2016: 301–302) speaks favourably of Bayesian methods because the Bayesian concept of probability is allegedly better-suited for cognitive modelling than the frequentist one. Her argument is part of a larger body of literature asking for cognitively plausible modelling techniques, for example Naïve Discriminative Learning (NDL; Baayen 2011; Baayen, Endresen, et al. 2013; Milin et al. 2016; Theijssen et al. 2013).[47] Yet, neither frequentist nor Bayesian methods were conceived of as cognitive models, but as systems of inference for scientists (see above, and see also Divjak 2016: 302). The fundamental question that lurks behind such arguments is how we interpret our statistical models (estimated on corpus data). Are they inductive models of cognitive representations which human learners would also infer through being exposed to the corpus data?[48] Or are they tests of theories that are pre-specified and merely tested for predictive accuracy on linguistic output data contained in corpora? In the former case, we adopt a strong *corpus as input* hypothesis (Stefanowitsch & Flach 2016) and should maybe resort to cognitively plausible statistical methods (whatever these might be). In the latter and less extreme case, the cognitive commitment does not necessarily extend to the statistical methods used. These methods, then, do not need to be any more cognitively plausible than an ANOVA used to analyse the results from an experiment in cognitive science. I view my own work in the tradition of theory testing, and cognitive realism is thus not a requirement for my methods of statistical inference of choice.

With regard to the third point, Levshina (2016: 251–252) argues for Bayesian estimation in mixed regression settings. First, she claims that "while frequentist statistics only allows one to test whether the null hypothesis can be rejected, Bayesian statistics enables one both to test the null hypothesis and to estimate the probability of specific parameter values given the data". This does not do justice to frequentist methods (and

[47] On p. 303 of Divjak (2016), the author goes on to explicitly mention NDL as well.

[48] In which case we would be doing *data science in language research* in the words of Milin et al. 2016. I see this as standing in contradiction to the view advocated in Dąbrowska (2016) as cited above (p. 14).

makes it sound like the author equates frequentism with NHST) in that mere rejection of the null hypothesis is characteristic only of Fisher's approach in its most rudimentary version. In the Neyman-Pearson approach, results actually *favour* one hypothesis over the other (cf. Lehmann 1993; 2011; Perezgonzalez 2015b) and lead to informed decision making. Furthermore, especially Neyman-style frequentism has well-known extensions to estimation, for example in the form of confidence intervals (see Greenland et al. 2016, esp. p. 340). Levshina then also explains that a "distinctive feature of Bayesian statistics is the use of so-called priors" and that "posterior probabilities depend on both the prior beliefs and the data, whereas the results of a frequentist model depend only on the data" (Levshina 2016: 252). Remarkably given this statement, she does *not* use informative priors, and in her footnote 8, Levshina (2016: 252) admits that priors were probed using trial and error. So, the proclaimed major advantage of Bayesian modeling was apparently not taken advantage of.[49]

Now, Maximum Likelihood Estimation (MLE) – the traditional method which could have been used instead of a Bayesian estimator – is not inherently *frequentist* in the sense of Neyman-Pearson testing theory. MLE, like inductive Bayesianism, conditions on the particular data inasmuch as it searches for the most likely set of parameters given the data. Frequentist testing theory is then used to make inferences based on variance parameters estimated by the ML estimator. What is more, Bayesian estimators are in fact based on the Likelihood and merely multiply it by the prior (Gelman, Carlin, et al. 2014: 6–8). If the prior is flat, results between MLE and Bayesian estimators converge (see also Gelman & Hill 2006: 347). The same is true if the sample size is large compared to the number of parameters, at least for finite-dimensional parameter models (Freedman 1999: 1119–1120), a well-established result known as the *Bernstein-von Mises theorem*. With a modest model structure including 17 fixed effects and 2,646 data points in Levshina (2016), it is highly likely that the same results would have been obtained with MLE. In fact, she admits that changing the priors did not lead to substantially different results in her footnote 8. This is a clear sign that the prior is "swamped by the data" (Freedman 1999: 1119). So far, I see no theoretically well-founded or practical arguments in favour of the Bayesian approach. If there had been evidence in Levshina's study that Bayesian and MLE methods did *not* converge, it would have been

[49] In the words of Senn (2011): "You may believe you are a Bayesian but you are probably wrong." Even Gelman & Hill (2006: 347–348) "view any noninformative prior distribution as inherently provisional" and give recommendations how to proceed once posteriors have been obtained from noninformative priors.

Level	Regressor	p_{PB}	Level	Coefficient		CI low		CI high		CI excludes 0	
				MLE	MCMC	MLE	MCMC	MLE	MCMC	MLE	MCMC
First	Badness	0.002		-0.152	-0.155	-0.247	-0.247	-0.061	-0.065	*	*
	Cardinal	0.001	No	1.189	1.222	0.862	0.927	1.466	1.496	*	*
	Genitives	0.001		-0.693	-0.711	-0.768	-0.801	-0.592	-0.616	*	*
	Measurecase	0.001	Acc	0.030	0.031	-0.150	-0.159	0.212	0.222		
			Dat	0.705	0.729	0.455	0.465	0.944	0.995	*	*
Second	Kindattraction	0.020		0.225	0.244	0.049	0.056	0.393	0.422	*	*
(Kind)	Kindfreq	0.095		0.146	0.164	-0.023	-0.016	0.301	0.341		
	Kindgender	0.001	Neut	0.021	0.013	-0.367	-0.409	0.392	0.435		
			Fem	1.269	1.289	0.800	0.788	1.709	1.783	*	*
Second	Measureattraction	0.001		0.282	0.299	0.106	0.102	0.447	0.515	*	*
(Measure)	Measureclass	0.001	Container	0.252	0.257	-0.265	-0.303	0.788	0.813		
			Rest	0.421	0.379	-0.209	-0.378	1.063	1.091		
			Amount	0.831	0.889	0.215	0.220	1.432	1.569	*	*
			Portion	1.217	1.253	0.675	0.689	1.684	1.840	*	*
	Measurefreq	0.005		-0.231	-0.232	-0.363	-0.395	-0.079	-0.073	*	*

Table 2: For the main study from Schäfer (2018): coefficient table comparing Maximum Likelihood Estimation (MLE, with 95% bootstrap confidence interval; 1,000 replications) and Bayesian Markov-Chain Monte Carlo estimation (MCMC; 4 chains; 1,000 iterations; normal priors for coefficients); the intercept (*Cardinal=Yes*, *Measurecase=Nom*, *Kindgender=Masc*, *Measureclass=Physical*; 0 for all numeric z-transformed regressors) is -3.548 (MLE) and -3.700 (MCMC).

an occasion to demonstrate the selective superiority of the algorithms used in Bayesian estimation. After all, there are situations where Bayesian estimators can be more robust, namely with heavily censored data, complex hierarchical models, perfect separation, etc. (see Freedman 1999, Gelman & Hill 2006: 345–348).

I want to state clearly that these points do not in any way invalidate the results presented in Levshina (2016). However, being "Bayesian" (as touted in the title of the paper) is most likely not among its selling points. Additionally, I want to voice the concern that many practitioners are probably already struggling with getting an adequate grasp of advanced statistical methods and that it might therefore be wise to use the more conservative and better understood method if the alternative method is not absolutely required for substantive reasons.

Finally, in order to demonstrate the convergence of the two types of estimators, I estimated the parameters of the hierarchical model presented in Schäfer (2018) with MLE and Markov-Chain Monte Carlo (MCMC) methods (the currently most prominent estimator used in Bayesian settings) to demonstrate their expectable convergence. Table 2 and the fixed effects coefficient plot in Figure 5 show the results.

This concludes the theoretical and methodological evaluation. I have defended the general approach to alternation modelling and probabilistic grammar, motivated the choice of data (mostly web corpora), and argued that the methods of statistical inference used in my research are indeed valid. In Section 3, I now put the four case studies into perspective before pointing to possible future research in Section 4.

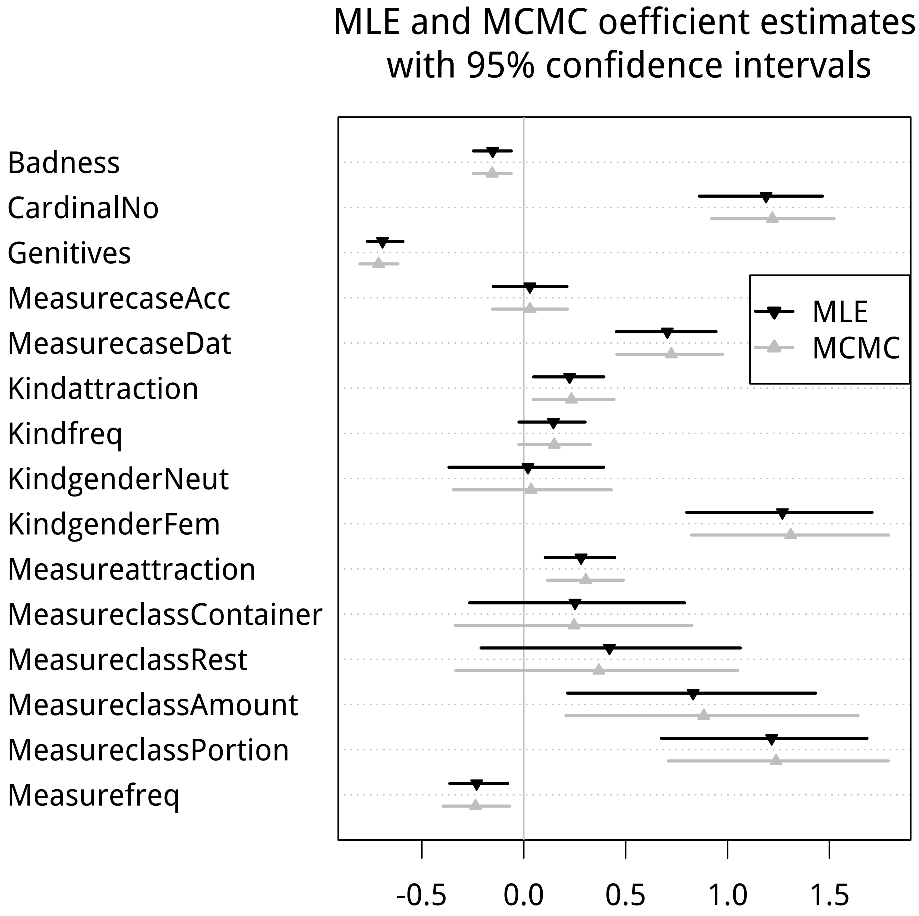

Figure 5: For the main study from Schäfer (2018): coefficient plot comparing Maximum Likelihood Estimation (MLE, with 95% bootstrap confidence interval; 1,000 replications) and Bayesian Markov-Chain Monte Carlo estimation (MCMC; 4 chains; 1,000 iterations; normal priors for coefficients); the intercept (*Cardinal=Yes*, *Measurecase=Nom*, *Kindgender=Masc*, *Measureclass=Physical*; 0 for all numeric z-transformed regressors) is -3.548 (MLE) and -3.700 (MCMC).

3 Case studies

In this section, I introduce the four case studies presented in detail in the published papers. I briefly describe the phenomenon under investigation in each paper, the assumed theoretical models, the methods used, and the paper's contribution to the research in probabilistic modelling and/or alternation modelling. In addition to these contributions, all papers represent innovation and advancement for the grammatical description of German, especially with a focus on non-standard written language and so-called *Zweifelsfälle* 'cases of doubt'. Should a large corpus-based descriptive grammar of German (which should obviously include alternations in its descriptions) ever be written, the studies presented here could serve as a blueprint for such a product. The details of the painstaking work (data collection, annotation, and fine-grained statistical analysis) reported in each individual paper demonstrate, however, that such an undertaking would be a monumental one lasting decades and requiring substantial manpower.

All studies used the DECOW web corpus (Schäfer & Bildhauer 2012; Schäfer 2015), which contains a mix of standard and non-standard written German as its main source of data (see Section 2.2.2). The statistics in each paper were programmed in the R programming language (R Core Team 2014).

Graphemic words and new paradigms: the cliticised indefinite article

In Schäfer & Sayatz (2014), we show how the emerging short forms of the German indefinite article create a new alternative paradigm, thus representing an independent alternative to the full forms. While the full standard forms use the stem *ein*, the stem is reduced to *n* in the short forms. The picture is complicated by forms like *nen*, which look like short forms of *einen* (accusative masculine singular), but which are also used as short forms corresponding to *ein* (nominative masculine/neuter singular). We argue that such forms have undergone a reconstruction process to fulfill both phonological and graphemic constraints on independent words. The paper reports five independent corpus studies. Study 1 shows that there is some amount of free variation, substantiated by the occurrence of a significant number of exemplars with one short form and one full form in NP conjunction structures with *und* 'and' or *oder* 'or' (regardless of the case of the NP and the order of the two conjuncts). Study 2 shows that the overall tendency of not using the genitive in non-standard documents alone does not account for the fact that the paradigm of the short forms does not have a genitive at all.[50] Study 3 presents a GLM which predicts the alternation between full and short forms using a number of theoretically motivated regressors, the main result being that the masculine and neuter nominative as well as the neuter accusative have the strongest tendency to preserve the full form, which was among the theoretically motivated predictions. Study 4 shows (using another, much simpler GLM) that the form *nen* is most likely preferred under certain morphophonological and graphemic conditions and not specifically marked for a morphosyntactic function. Finally, Study 5 shows that independent graphemic principles guide writers' behaviour, as full graphemic enclisis (contraction of the reduced indefinite article with the preceding word without space or apostrophe) occurs predominantly with the form *n*. This, we argue, can only be explained by the fact that *n* is not at all a prototypical independent graphemic word.

This paper is written in a descriptive tone, not making claims about cognitive representations. However, since general principles (such as universal and language-specific conditions on enclisis) are referred to, a cognitive interpretation and an experimental cross-validation would be possible. The paper fits well into the alternation research paradigm, as other studies have been presented within it which also model choices between full forms and contracted (or cliticised) forms, e. g. Barth & Kapatsinski (2014). We explicitly argue for the benefits of using web corpora (DECOW12 in this case), because the short forms are entirely absent from standard written language.

Prototypes and paradigms: the strength of weak nouns

In Schäfer (2016c), I demonstrate how robust corpus-based models can be constructed and verified based on previous substantive prototype-theoretical research. The approximately five hundred masculine weak nouns in German like *Mensch* 'human, man' follow a remarkably odd inflectional paradigm compared to all other nouns. They mark all forms except for the nominative singular with *-en*. While there is another paradigm of masculine/neuter nouns which mark the plural with *-en*, the weak singular forms in the accusative, dative, and genitive are truly exceptional. Furthermore, there is a non-standard alternation inasmuch as weak nouns sometimes occur in strong forms. If that happens, they simply drop the *-en* in the accusative and dative, and they take on the typical strong ending *-es* instead of *-en* in the genitive. From Thieroff (2003) – an analysis of the relevant paradigm structure – I predict that the strong forms of weak nouns should

[50] The expected genitive forms *nes* and *ner* virtually do not occur.

be more frequent in the accusative and dative than in the genitive. Base on the prototype-theoretical analysis in Köpcke (1995), I also predict that the strong forms should be less frequent when the nouns denote humans and when certain phonotactic conditions are met.[51] The predictions are borne out in a large-scale corpus analysis using a GLM to predict the alternants.

Obviously, the paper fits well into the alternation modelling approach. It uses a cognitively motivated model (based on prototype theory) and all the standard tools described in Section 2. In the future, a re-analysis using a per-lemma random effect will be attempted. As it stands, lemma-specific effects are not modelled because the estimator did not converge with random effects in the model. Since new algorithms and statistical packages are constantly being made available, a more realistic statistical model might be possible in the future.

Prototypes and grammaticalisation: the measure NP alternation

In Schäfer (2018), I model a case alternation in German measure noun phrases such as *ein Fass reines Öl* (both nouns have identical case) or *ein Fass reinen Öls* (the kind-denoting noun has genitive case), both 'a barrel of pure oil'. I describe the prototypical meanings of both alternants in terms of the grammaticalisation paths leading to partitive and pseudo-partitive (Koptjevskaja-Tamm 2001) constructions. Basically, the genitive construction is assumed to prototypically allow both a referent of the measure noun and of the kind-denoting noun to be accessible. The same is possible but less prototypical for the case identity construction. From the definition of the prototypes, a number of predictions are derived regarding the preferences of measure nouns from different semantic classes to occur in one alternant or the other. Also, a prediction regarding cardinal or non-cardinal determiners is derived from the prototypical meanings.

Furthermore, an exemplar effect is modelled. For the alternating construction, there exist two neighbouring construction which always require the genitive ('ein Fass des reinen Öls' with a determiner on the kind-denoting noun) or never allow the genitive ('ein Fass Öl' with no determiner and no adjective). The occurrence frequencies of measure and kind lemmas in these two constructions is shown to influence the alternation in the expected direction.

[51] Koepcke shows (among other things, based on diachronic data) that the weak nouns predominantly represent a semantically and phonotactically well-defined prototype. The features enumerated in his paper and used in my model are slightly more fine-grained than this short introduction makes them sound.

The study is a prime example of corpus-based alternation research based on substantive theory. Not only is each predictor in the multilevel model independently motivated, but there is also experimental cross-validation in two experimental paradigms (forced choice and self-paced reading). The paper makes significant contributions to the prototype vs. exemplar debate, and it discusses the question of how well corpus-derived models and experimental validations can be expected to converge.

Prototypical syntax and punctuation: (non-)embedded V2 clauses

Finally, in Schäfer & Sayatz (2016), Ulrike Sayatz and I use graphemic data from non-standard written German to substantiate conclusions about sentential structure in embedded and non-embedded clauses introduced by the particles *obwohl* 'although, then again' and *weil* 'because'. Both particles are subordinators in standard written German and as such embed clauses with verb-last constituent order. It has been known for quite a while that with some semantic and pragmatic changes they can also embed verb-second clauses, which is the constituent order otherwise typical of independent clauses.

We perform an in-depth analysis of the punctuation occurring before and after the two particles, also using GLMs. The results are very strong and could even have been detected with descriptive statistics alone. It turns out that *obwohl* with verb-second order occurs proportionally more often at the beginning of sentences after full stops. Also, *obwohl* is separated more often from the clause it embeds by punctuation marks which are otherwise used with sentence-initial, verb-second-embedding discourse particles such as *natürlich* 'naturally' of *klar* 'of course'. In line with previous research, we argue that the distribution of the punctuation marks provides solid evidence that the two particles (with verb-second order) have different syntactic and pragmatic functions and that *obwohl* is essentially a discourse particle used in independent sentences.

The paper does not straightforwardly belong into the alternation research category, but it clearly models a probabilistic phenomenon, as the syntactic structures, the pragmatic functions, and the graphemic markers are subject to stochastic variation. The major contribution of the paper is the first-ever proposal of *usage-based graphemics* (UBG). We understand UBG neither as a theory nor as a framework. Rather, we see it as a method of analysing graphemic variation as a cue to grammatical structure. It is a method of analysis which has the potential to develop into a framework concerned with the syntax-graphemics interface.

We assume that writers learn to associate phonological, morphological, and syntactic patterns with graphemic patterns through repeated exposure to the graphemic patterns in conjunction with the grammatical ones.[52] Thus, they learn to associate grammatical prototypes (such as sentence type prototypes) probabilistically with graphemic units and patterns such as punctuation marks. This is clearly in line with assumption of usage-based theories (Bybee & Beckner 2009). Especially when the normative pressure is low (as is the case when, for example, German writers start using the completely non-standard *obwohl* and *weil* clauses with verb-second order in writing) and writers have to encode syntactic structures which are novel or usually not encoded at all in writing, the prototypical mapping becomes visible through emerging regularities where no normative rules exist.

What is promising about this approach is that – once it is fully developed – it can be used to reconstruct evidence for grammatical structure from corpora containing non-standard writing. For this to work reliably, the correspondences and mechanisms have to be fleshed out, and experimental validation is required. Therefore, it was vital that the results converged with previous analyses of the two particles, thus substantiating the assumption that UBG is a valid method of analysis. UBG will be described in more detail in Schäfer & Sayatz (n.d.) and several other publications by Ulrike Sayatz and me which include experimental work.

[52] Despite some construction terminology used in our paper, UBG is not necessarily tied to construction grammar or any other grammatical framework. Any system of units of grammar and their combinatorics can be mapped onto graphemic patterns.

4 Future directions

My and my co-author's research collected here shows that German has a wide range of phenomena to offer for examination under a usage-based probabilistic perspective. Furthermore, in the form of the DECOW corpus, a now-proven source of data exists which allows researchers to work on these phenomena. Based on the argumentation in Sections 1–3 and the case studies, a number of open research questions come to mind. I see at least the following ones.

- The literature on *cases of doubt* in German is famously rich. It would be beneficial for linguists working on German, corpus linguists, and linguists working in the cognitively oriented/usage-based tradition to examine them using the framework established here.
- The effects of corpus composition and the availability of metadata on corpus samples and sampling procedures should be examined further. The BNC is rich in metadata and has a well-planned composition, but for many other languages (like German), similar corpora do not exist.
- Related to the last point, corpora containing non-standard writing should be honoured more as a unique source of data. While there is a community working on such corpora and specific analyses of their content, many more (corpus) linguists could benefit from using them in the same way I did.
- Also related to this point, the usage-based perspective on graphemics as developed by Ulrike Sayatz and me should be developed and expanded further. Speakers' writing behaviour provides important clues to how they cognitively represent morphological and syntactic categories.
- Individual grammatical differences urgently require more attention. While it will probably be impossible to build large enough general-purpose corpora with speaker metadata which would allow research on individual grammatical differences, corpus data should be correlated with the reactions of individual speakers in controlled experiments.
- The prototype vs. exemplar debate would benefit from more large-scale corpus studies which must then be cross-validated in controlled experiments. Corpus data alone cannot provide evidence for or against one theory or the other.

- Statistical methods need to be scrutinised. While mindless applications of NHST are detrimental for valid scientific inferences, some critiques of frequentist statistics (language is never random; model everything) have gone too far or are understood in a much too unrestricted manner. Also, some currently-hyped alternative methods do not lead to substantially different results, are understood even less than traditional methods, and distract from the real problems with statistical inference.

Clearly, my work has contributed to all of these points, but the overall situation in probabilistic, usage-based, cognitively oriented corpus linguistics is one where methods and theories are still in a very early stage of development.

References

Anderson, Kenneth P. & David R. Burnham. 2002. *Model selection and multimodel inference: a practical information-theoretic approach*. Berlin: Springer.

Arnold, Doug & Evita Linardaki. 2007. HPSG-DOP: Towards exemplar-based HPSG. In *Proceedings of the workshop on exemplar based models of language acquisition and use*, 42–51. Dublin.

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.

Baayen, R. Harald. 2011. Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics* 11. 295–328.

Baayen, R. Harald, Anna Endresen, Laura A. Janda, Anastasia Makarova & Tore Nesset. 2013. Making choices in russian: pros and cons of statistical methods for rival forms. *Russian Linguistics* 37. 253–291.

Baayen, R. Harald, Cyrus Shaoul, Jon Willits & Michael Ramscar. 2016. Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience* 31(1). 106–128.

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.

Barr, Dale J., Roger Levy, Christoph Scheepers & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language* 68(3). 255–278.

Barsalou, Lawrence W. 1990. On the indistinguishability of exemplar memory and abstraction in category representation. In Thomas K. Srull & Robert S. Wyer (eds.), *Advances in social cognition, volume III: content and process specificity in the effects of prior experiences*, 61–88. Hillsdale: Erlbaum.

Barth, Danielle & Vsevolod Kapatsinski. 2014. 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* ahead of print.

Bates, Douglas M. 2010. Lme4: mixed-effects modeling with R.

Bates, Douglas M., Reinhold Kliegl, Shravan Vasishth & R. Harald Baayen. 2015. *Parsimonious Mixed Models*.

Berk, Richard A. & David A. Freedman. 2009. Statistical assumptions as empirical commitments. In David Collier, Jasjeet S. Sekhon & Philip B.

Stark (eds.), *Statistical models and causal inference: a dialogue with the social sciences*, 23–44. Cambridge: Cambridge University Press.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8(4). 243–257.

Biber, Douglas & Jesse Egbert. 2016. Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. *Journal of Research Design and Statistics in Linguistics and Communication Science* 2. 3–36.

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.

Bildhauer, Felix & Roland Schäfer. 2016. Automatic classification by topic domain for meta data generation, web corpus evaluation, and corpus comparison. In Paul Cook, Stefan Evert, Roland Schäfer & Egon Stemle (eds.), *Proceedings of the 10th web as corpus workshop (WAC-X)*, 1–6. Berlin: Association for Computational Linguistics.

Bildhauer, Felix & Roland Schäfer. 2017. Induktive Topikmodellierung und extrinsische Topikdomänen. In Marek Konopka & Angelika Wöllstein (eds.), *Grammatische Variation - empirische Zugänge und theoretische Modellierung* (Jahrbuch des Instituts für Deutsche Sprache 2016), 331–344. Berlin & Boston: De Gruyter.

Birnbaum, Allan. 1962. On the foundations of statistical inference. *Journal of the American Statistical Association* 57(298). 269–326.

Bod, Rens. 2006. Exemplar-based syntax: how to get productivity from examples. *The Linguistic Review* 23. 291–320.

Bortz, Jürgen. 2005. *Statistik für Human- und Sozialwissenschaftler*. 6th edn. Heidelberg: Springer.

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* (Studies in Generative Grammar), 77–96. Berlin/New York: De Gruyter Mouton.

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.

Burnard, Lou. 2007. *The British National Corpus Users Reference Guide*. Tech. rep. University of Oxford.

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.

Campbell, Donald T. & Donald W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56. 81–105.

Chalmers, Alan. 2013. *What is this thing called science*. 4th edn. Maidenhead: McGraw Hill.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Colquhoun, David. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* 1(3). 140216.

Conaway, Nolan & Kenneth J. Kurtz. 2016. Similar to the category, but not the exemplars: a study of generalization. *Psychonomic Bulletin & Review* 24. 1312–1323.

Cronbach, Lee J. & Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52. 281–302.

Cumming, Geoff. 2014. The new statistics: why and how. *Psychological Science* 25(1). 7–29.

Dąbrowska, Ewa. 2008. The effects of frequency and neighbourhood density on adult speakers' productivity with Polish case inflections: an empirical test of usage-based approaches to morphology. *Journal of Memory and Language* 58. 931–951.

Dąbrowska, Ewa. 2012. Different speakers, different grammars: individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2. 219–253.

Dąbrowska, Ewa. 2014. Words that go together: measuring individual differences in native speakers' knowledge of collocations. *The Mental Lexicon* 9(3). 401–418.

Dąbrowska, Ewa. 2015. Individual differences in grammatical knowledge. In Ewa Dąbrowska & Dagmar Divjak (eds.), *Handbook of cognitive linguistics*, 650–668. De Gruyter.

Dąbrowska, Ewa. 2016. Cognitive linguistics' seven deadly sins. *Cognitive Linguistics* 27(4). 479–491.

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.

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

Divjak, Dagmar, Antti Arppe & R. Harald Baayen. 2016. Does language-as-used fit a self-paced reading paradigm? In Tanja Anstatt, Anja Gattnar & Christina Clasmeier (eds.), *Slavic languages in psycholinguistics,* 52–82. Tübingen: Narr Francke Attempto.

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.

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

Duden. 2011. *Richtiges und gutes Deutsch - Das Wörterbuch der sprachlichen Zweifelsfälle*. Dudenredaktion unter Mitarbeit von Peter Eisenberg und Jan Georg Schneider (ed.). 7th edn. Mannheim/Zürich: Dudenverlag.

Durrant, Philip & Alice Doherty. 2010. Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory* 6(2). 125–155.

Elman, Jeffrey L. 2009. On the meaning of words and dinosaur bones: lexical knowledge without a lexicon. *Cognitive Science* 33. 547–582.

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

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang & Brian Marx. 2013. *Regression – models, methods, and application*. Berlin etc.: Springer.

Fisher, Ronald A. 1926. The arrangement of field experiment. *Journal of the Ministry of Agriculture of Great Britain* 33. 503–513.

Fodor, Janet D. 1995. Thematic roles and modularity. In Garry T. M. Altmann (ed.), *Cognitive models of speech processing,* 434–456. Cambridge: MIT Press.

Fox, John. 2016. *Applied regression analysis & generalized linear models*. 3rd edn. London: Sage Publications.

Freedman, David A. 1999. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics* 27(4). 1119–1140.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari & Donald B. Rubin. 2014. *Bayesian data analysis*. 3rd edn. Boca Raton: Chapman & Hall.

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Gelman, Andrew & Cosma Rohilla Shalizi. 2013. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* 66. 8–38.

Gigerenzer, Gerd. 2004. Mindless statistics. *The Journal of Socio-Economics* 33. 587–606.

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.

Good, Phillip I. & James W. Hardin. 2012. *Common errors in statistics (and how to avoid them)*. 4th edn. Hoboken: Wiley & Sons.

Goodman, Steven. 2008. A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology*. 135–140.

Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman & Douglas G. Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31. 337–350.

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. 2012. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics – some necessary clarifications. *Studies in Language* 11(3). 477–510.

Gries, Stefan Th. 2015a. More (old and new) misunderstandings of collostructional analysis: on Schmid and Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536.

Gries, Stefan Th. 2015b. The most underused statistical method in corpus linguistics: multi-level (and mixed-effects) models. *Corpora* 10(1). 95–126.

Gries, Stefan Th. 2017a. Corpus approaches. In Barbara Dancygier (ed.), *Cambridge handbook of cognitive linguistics*, 590–606. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2017b. Syntactic alternation research. taking stock and some suggestions for the future. In Ludovic De Cuypere, Clara Vanderschueren & Gert De Sutter (eds.), *Current trends in analyzing syntactic variation*, vol. 31 (Belgian Journal of Lingustics), 7–27. Amsterdam: Benjamins.

Gries, Stefan Th., Beate Hampe & Doris Schönefeld. 2005. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676.

Gries, Stefan Th. & Anatol Stefanowitsch. 2004. Co-varying collexemes in the into-causative. In Michel Achard & Suzanne Kemmer (eds.), *Language, culture, and mind*, 225–236. Stanford, CA: CSLI.

Gries, Stefan Th. & Stefanie Wulff. 2005. Do foreign language learners also have constructions? evidence from priming, sorting, and corpora. *Annual Review of Cognitive Linguistics* 3. 182–200.

Griffiths, Thomas L., Kevin R. Canini, Adam N. Sanborn & Daniel J. Navarro. 2009. Unifying rational models of categorization via the hierarchical dirichlet process. In *Proceedings of the 29th annual conference of the cognitive science society*, 323–328. Mahwah: Erlbaum.

Hahn, Ulrike, Mercè Prat-Sala, Emmanuel M. Pothos & Duncan P. Brumby. 2010. Exemplar similarity and rule application. *Cognition* 114(1). 1–18.

Harris, Zellig. 1954. Distributional structure. *Word* 10(23). 146–162.

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

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

Huettig, Falk & Esther Janse. 2016. Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience* 31(1). 80–93.

Hunston, Susan. 2008. Collection strategies and design decisions. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook*, 154–168. Berlin: Walter de Gruyter.

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

Kilgarriff, Adam. 2005. Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2). 263–276.

Kilgarriff, Adam. 2006. Googleology is bad science. *Computational Linguistics* 33(1). 147–151.

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.

Klein, Wolf-Peter. 2009. Auf der Kippe? Zweifelsfälle als Herausforderung(en) für Sprachwissenschaft und Sprachnormierung. In Marek Konopka & Bruno Strecker (eds.), *Deutsche Grammatik – Regeln, Normen, Sprachgebrauch*. Berlin: De Gruyter.

Köpcke, Klaus-Michael. 1995. Die Klassifikation der schwachen Maskulina in der deutschen Gegenwartssprache – Ein Beispiel für die Leistungsfähigkeit der Prototypentheorie. *Zeitschrift für Sprachwissenschaft* 14(2). 159–180.

Koptjevskaja-Tamm, Maria. 2001. "A piece of the cake" and "a cup of tea": partitive and pseudo-partitive nominal constructions in the Circum-Baltic languages. In Östen Dahl & Maria Koptjevskaja-Tamm (eds.), *The Circum-Baltic languages: typology and contact*, vol. 2, 523–568. Amsterdam & Philadelphia: John Benjamins.

Krause, Anne. 2016. Proceedings of the 10th Web as Corpus Workshop. In. Berlin: Association for Computational Linguistics. Chap. The Challenges and Joys of Analysing Ongoing Language Change in Web-based Corpora: a Case Study, 27–34.

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.

Kuhn, Thomas. 1970. *The structure of scientific revolutions*. 2nd edn. Chicago: University of Chicago Press.

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.

Kupietz, Marc, Cyril Belica, Holger Keibel & Andreas Witt. 2010. The German reference corpus DeReKo: a primordial sample for linguistic research. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias (eds.), *Proceedings of the seventh international conference on language resources and evaluation (LREC '10)*, 1848–1854. Valletta, Malta: European Language Resources Association (ELRA).

Langacker, Ronald W. 1987. *Foundations of cognitive grammar (volume 1: theoretical prerequisites)*. Stanford: Stanford University Press.

Lee, David. 2001. Genres, registers, text types, domains, and styles: claryfying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3). 37–72.

Lee, Michael D. & Wolf Vanpaemel. 2008. Exemplars, prototypes, similarities, and rules in category representation: an example of hierarchical Bayesian analysis. *Cognitive Science* 32. 1403–1424.

Leech, Geoffrey. 2007. New resources or just better old ones? The Holy Grail of representativeness. In Marianne Hundt, Nadja Nesselhauf & Carolin Biewer (eds.), *Corpus linguistics and the web*, 133–149. Amsterdam & New York: Rodopi.

Lehmann, Erich L. 1993. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistics Association* 88. 1242–1249.

Lehmann, Erich L. 2011. *Fisher, Neyman, and the creation of classical statistics*. New York, NY: Springer.

Levshina, Natalia. 2016. When variables align: a Bayesian multinomial mixed-effects model of English permissive constructions. *Cognitive Linguistics* 27(2). 235–268.

Manning, Christopher D. 2002. Probabilistic syntax. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 289–341. Cambridge: MIT Press.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen & Douglas M. Bates. 2017. Balancing type I error and power in linear mixed models. *Journal of Memory and Language* 94. 305–315.

Maxwell, Scott E. & Harold D. Delaney. 2004. *Designing experiments and analyzing data: a model comparison perspective*. Mahwa, New Jersey, London: Taylor & Francis.

Mayo, Deborah G. 1996. *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G. (ed.). 2009. *Error and inference: recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science*. Cambridge: Cambridge University Press.

Mayo, Deborah G. 2011. How can we cultivate Senn's ability? *Rationality, Markets, and Morals* 3. 14–18.

Mayo, Deborah G. 2013. The error-statistical philosophy and the practice of Bayesian statistics: comments on Gelman and Shalizi: 'Philosophy and the practice of Bayesian statistics'. *British Journal of Mathematical and Statistical Psychology* 66. 57–64.

Mayo, Deborah G. 2014. On the Birnbaum argument for the strong likelihood principle. *Statistical Science* 29. 227–266.

Mayo, Deborah G. 2018. *Statistical inference as severe testing: how to get beyond the statistics wars*. Cambridge: Cambridge University Press.

Mayo, Deborah G. & Aris Spanos. 2006. Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science* 57. 323–357.

McEnery, Tony, Richard Xiao & Yukio Tono. 2006. *Corpus-based language studies. An advanced resource book*. London & New York: Routledge.

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

Milin, Petar, Dagmar Divjak, Strahinja Dimitrijević & R. Harald Baayen. 2016. Towards cognitively plausible data science in language research. *Cognitive Linguistics* 27(4). 507–526.

Minda, John Paul & J. David Smith. 2001. Prototypes in category learning: the effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27(3). 775–799.

Minda, John Paul & J. David Smith. 2002. Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28(2). 275–292.

Mollin, Sandra. 2009. Combining corpus linguistic and psychological data on word co-occurrences: corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5(2). 175–200.

Müller, Stefan. 2018. *Grammatical theory: From Transformational Grammar to constraint-based approaches*. 2nd edn. (Textbooks in Language Sciences 1). Berlin: Language Science Press.

Munafò, Marcus R., Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware & John P. A. Ioannidis. 2017. A manifesto for reproducible science. *Nature Human Behaviour*. 0021 EP.

Murphy, Gregory. 2002. *The big book of concepts*. Cambridge: MIT Press.

Murphy, Gregory L. 2003. Ecological validity and the study of concepts. In Brian H. Ross (ed.), *Psychology of learning and motivation - advances in research and theory*, 1–41. New York: Elsevier.

Nesset, Tore & Laura A. Janda. 2010. Paradigm structure: evidence from Russian suffix shift. *Cognitive Linguistics* 21(4). 699–725.

Newman, John & Tamara Sorenson Duncan. 2015. *Convergence and divergence in Cognitive Linguistics: Facing up to alternative realities of linguistic catgeories*. Talk given at the 13th international cognitive linguistics conference (ICLC-13).

Newmeyer, Frederick J. 2003. Grammar is grammar and usage is usage. *Language* 79(4). 682–707.

Neyman, Jerzy. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society A* 236(767). 333–379.

Neyman, Jerzy & Egon S. Pearson. 1933. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A* 231. 694–706.

Partee, Barbara, Alice ter Meulen & Robert E. Wall. 1990. *Mathematical methods in linguistics*. Dordrecht: Kluwer.

Perezgonzalez, Jose D. 2014. A reconceptualization of significance testing. *Theory & Psychology* 24(6). 852–859.

Perezgonzalez, Jose D. 2015a. Confidence intervals and tests are two sides of the same research question. *Frontiers in Psychology*.

Perezgonzalez, Jose D. 2015b. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology* 6(223). 1–11.

Posner, Michael I. & Steven W. Keele. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology* 77(3). 353–363.

Pullum, Geoffrey K. 2013a. Consigning phenomena to performance: a response to neeleman. *Mind & Language* 28(4). 532–537.

Pullum, Geoffrey K. 2013b. The central question in comparative syntactic metatheory. *Mind & Language* 28(4). 492–521.

R Core Team. 2014. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Ramscar, Michael & Robert F. Port. 2016. How spoken languages work in the absence of an inventory of discrete units. *Language Sciences* 53. 58–74.

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.

Rosseel, Yves. 2002. Mixture models of categorization. *Journal of Mathematical Psychology* 46(2). 178–210.

Schäfer, Roland & Ulrike Sayatz. N.d. Gebrauchsbasierte Graphematik des Deutschen. in preparation.

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

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.

Schäfer, Roland. N.d. Statische inferenz in der linguistik. in preparation.

Schäfer, Roland. 2010. *Arguments and adjuncts at the syntax-semantics interface*. Göttingen: Georg-August Universität Dissertation zur Erlangung des Doktorgrades (Dr. phil.) im Fach Englische Philologie.

Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lüngen & Andreas Witt (eds.), *Proceedings of challenges in the management of large corpora 3 (CMLC-3)*. Lancaster: IDS.

Schäfer, Roland. 2016a. CommonCOW: massively huge web corpora from CommonCrawl data and a method to distribute them freely under restrictive EU copyright laws. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*, 4500–4504. Portorož: European Language Resources Association (ELRA).

Schäfer, Roland. 2016b. *Einführung in die grammatische Beschreibung des Deutschen*. 2nd edn. Berlin: Language Science Press.

Schäfer, Roland. 2016c. Prototype-driven alternations: the case of German weak nouns. *Corpus Linguistics and Linguistic Theory* ahead of print.

Schäfer, Roland. 2017. Accurate and efficient general-purpose boilerplate detection for crawled web corpora. *Language Resources and Evaluation* 51. 873–889.

Schäfer, Roland. 2018. Abstractions and exemplars: the measure noun phrase alternation in German. *Cognitive Linguistics* 29(4). 729–771.

Schäfer, Roland. 2019. Generalized linear mixed models. In Stefan Gries & Magali Paquot (eds.)*, The practical handbook of corpus linguistics*. in press. Berlin, Heidelberg: Springer.

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 (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 & Felix Bildhauer. 2013. *Web corpus construction* (Synthesis Lectures on Human Language Technologies). San Francisco: Morgan & Claypool.

Schäfer, Roland & Elizabeth Pankratz. 2018. The plural interpretability of German linking elements. *Morphology* 28(4). 325–358.

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.

Senn, Stepen J. 2011. You may believe you are a Bayesian but you are probably wrong. *Rationality, Markets, and Morals* 2. 48–66.

Sprouse, Jon & Diogo Almeida. 2012. Assessing the reliability of textbook data in syntax: Adger's Core Syntax. *Journal of Linguistics* 48. 609–652.

Sprouse, Jon, Carson T. Schütze & Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010. *Lingua* 134. 219–248.

Stefanowitsch, Anatol & Susanne Flach. 2016. A corpus-based perspective on entrenchment. In Hans-Jörg Schmid (ed.), *Entrenchment, memory and automaticity. The psychology of linguistic knowledge and language learning*, 101–128. Berlin: De Gruyter.

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

Storms, Gert, Paul De Boeck & Wim Ruts. 2000. Prototype and exemplar-based information in natural language categories. *Journal of Memory and Language* 42. 51–73.

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.

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.

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen & hans van Halteren. 2013. Choosing alternatives: using Bayesian networks and memory-based learning to study the dative alternation. *Corpus Linguistics and Linguistic Theory* 9(2). 227–262.

Thieroff, Rolf. 2003. Die bedienung des automatens durch den mensch. deklination der schwachen maskulina als zweifelsfall. *Linguistik Online* 16(4). 105–117.

Trafimow, David & Michael Marks. 2015. Editorial. *Basic and Applied Social Psychology* 37(1). 1–2.

Tummers, José, Kris Heylen & Dirk Geeraerts. 2005. Usage-based approaches in cognitive linguistics: a technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2). 225–261.

Vanpaemel, Wolf. 2016. Prototypes, exemplars and the response scaling parameter: a Bayes factor perspective. *Journal of Mathematical Psychology* 72. 183–190.

Vanpaemel, Wolf & Gert Storms. 2008. In search of abstraction: the varying abstraction model of categorization. *Psychonomic Bulletin & Review* 15(4). 732–749.

Verbeemen, Timothy, Wolf Vanpaemel, Sven Pattyn, Gert Storms & Tom Verguts. 2007. Beyond exemplars and prototypes as memory representations of natural concepts: a clustering approach. *Journal of Memory and Language* 56(4). 537–554.

Vigen, Tyler. 2015. *Spurious correlations*. New York: Hachette.

Voorspoels, Wouter, Wolf Vanpaemel & Gert Storms. 2011. A formal ideal-based account of typicality. *Psychonomic Bulletin & Review* 18. 1006–1014.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. 2nd edn. Cambridge: MIT Press.

Wulff, Stefanie. 2003. A multifactorial corpus analysis of adjective order in english. *International Journal of Corpus Linguistics* 8(2). 245–282.

Zimmer, Christian. 2015. Bei einem Glas guten Wein(es): Der Abbau des partitiven Genitivs und seine Reflexe im Gegenwartsdeutschen. *Beiträge zur Geschichte der deutschen Sprache und Literatur* 137(1). 1–41.

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.