GRAMMAR AND CORPORA 2016

edited by
Eric Fuß
Marek Konopka
Beata Trawiński
Ulrich H. Waßner

HEIDELBERG
UNIVERSITY PUBLISHING

Grammar and Corpora

2016

Grammar and Corpora 2016

edited by

Eric Fuß
Marek Konopka
Beata Trawiński
Ulrich H. Waßner

HEIDELBERG
UNIVERSITY PUBLISHING

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie. Detailed bibliographic data are available on the Internet at http://dnb.dnb.de.

The electronic, open access version of this work is permanently available on Heidelberg University Publishing's website: http://heiup.uni-heidelberg.de.
urn: urn:nbn:de:bsz:16-heiup-book-361-6
doi: https://doi.org/10.17885/heiup.361.509

ISBN 978-3-946054-82-5 (Softcover)
ISBN 978-3-946054-83-2 (Hardcover)
ISBN 978-3-946054-84-9 (PDF)

Contents

II. Methodology and Application

Current Trends and Issues

Epilogue

Prologue

Eric Fuß, Marek Konopka, Beata Trawiński, Ulrich H. Waßner

Grammar and Corpora – Past, Present, and Future

In recent years, the availability of large annotated and searchable corpora, together with a new interest in the empirical foundation and validation of linguistic theory and description, has sparked a surge of novel and interesting work using corpus-based methods to study the grammar of natural languages. However, a look at relevant current research on the grammar of the Germanic, Romance, and Slavic languages reveals a variety of different theoretical approaches and empirical foci, which can be traced back to different philological and linguistic traditions. Still, this current state of affairs should not be seen as an obstacle but as an ideal basis for a fruitful exchange of ideas between different research paradigms.

Starting from this premise, the sixth international conference *Grammar and Corpora*, of which the present volume is a result, took place at the Institut für Deutsche Sprache (IDS, Institute for the German Language) in Mannheim, Germany, from the 9th to the 11th of November 2016. The *Grammar and Corpora* conference series was founded by František Štícha (Academy of Sciences of the Czech Republic) in Prague in 2005.[1] While the first conference was largely devoted to corpus-oriented projects in the field of Slavic linguistics (mainly Czech), the programme of the second gathering in Liblice, Czech Republic, in 2007[2] already included research on other languages and methodological cross-linguistic perspectives. When Mannheim hosted the third conference in 2009,[3] the number of contributions on Germanic and Romance languages increased significantly.

1 Cf. Štícha and Šimandl (2007).
2 Cf. Štícha and Fried (2008).
3 Cf. Konopka et al. (2011).

After the conferences in Prague (2012)[4] and Warsaw (2014),[5] organised by the Czech Academy of Sciences and the Polish Academy of Sciences respectively, Mannheim became the venue for the second time. In 2016 the IDS welcomed 120 attendees who represented over 40 institutions from 16 countries. The conference was comprised of 35 regular papers and 15 poster presentations devoted to corpus-oriented projects focusing on Germanic, Slavic, and Romance languages, as well as to cross-linguistic methodology.

The internationalisation of the conference series reflects the fact that the field of corpus linguistics has always been a global enterprise, in which researchers from different countries collaborate. This is mainly because of the need to keep up with the methodological development of corpus collection, annotation, and analysis worldwide. This development builds upon the increasing availability of powerful computers that less and less often stops at country borders. Thus, although the study of individual languages was given center stage, cross-linguistic aspects have always played an important role in corpus-oriented grammar research. More generally, the development of the conference series mirrors the growing importance of linguistic research based on corpora over the last 30 years, which has been fueled by the need for a more solid empirical foundation of linguistic theory. Linguistics needs linguistic data, and corpora can provide huge amounts of data – much more data than introspections, interviews, questionnaires, or experiments. Moreover, contrary to the other empirical approaches, corpora usually provide authentic and spontaneous data that have not been induced by a researcher. Taking all this into account, the promotion of the use of corpus linguistic methods in research on grammar has been a major goal of all six conferences up to now. Accordingly, the conferences had to introduce methodological innovations and explore their potential uses in investigations of as wide a range of grammatical topics as possible. Therefore the only thematic limitation on the contributions (apart from the focus on certain languages) was that they had to combine work on grammar with an examination of corpus data.

Indeed, the papers and poster presentations of *Grammar and Corpora 2016* addressed a wide array of issues and covered different domains of linguistic analysis including phonology, morphology, syntax, text linguistics, and application-oriented studies. In addition, the conference attendees discussed and became acquainted with different methodological approaches, including more traditional methods as well as recent statistical and computer-linguistic based techniques and procedures.

4 Cf. <http://www.ujc.cas.cz/veda-vyzkum/vyzkum/gramatika-a-korpus/proceedings-2012/proceedings-gac-2012.html> (7.5.2018).

5 Cf. <http://ispan.waw.pl/default/images/konferencje/2014/gramatyka_korpus.pdf> (7.5.2018).

For the first time in the history of the *Grammar and Corpora* conference series, the 2016 conference was preceded by a Tutorial Day. The aim of this one-day, partly two-track tutorial programme was to provide a theoretical background and practical instructions on selected resources and applications related to the topics of the conference. It was comprised of four tutorials:

— "Working with Web Corpora" by Felix Bildhauer (IDS Mannheim) and Roland Schäfer (Freie Universität Berlin), cf. Schäfer (2015, 2016) and <http://corpora fromtheweb.org/> (7.5.2018)
— "InterCorp: Exploring a Multilingual Parallel Corpus" by Alexandr Rosen (Charles University Prague), cf. Čermák/Rosen (2012) and <https://wiki.korpus.cz/doku.php/en:cnk:intercorp> (7.5.2018)
— "Visualisierung linguistischer Daten mit der freien Grafik- und Statistikumgebung R" by Sandra Hansen-Morath and Sascha Wolfer (IDS Mannheim), cf. Hansen-Morath/Wolfer (2017) and <http://kograno.ids-mannheim.de/VisR-OnlinePub/> (7.5.2018)
— "Introduction to Corpus Analysis with KorAP" by Nils Diewald and Eliza Margaretha (IDS Mannheim), cf. Kupietz et al. (2017) and <http://korap.ids-mannheim.de/> (7.5.2018)

An overview of the tutorial day is available at the conference homepage under <http://gac2016.ids-mannheim.de> (7.5.2018). In addition, a report about the entire event is given (in German) by Münzberg (2016).

It should be noted that the content of the present volume is not identical to the conference programme. Rather, in preparing the collection at hand, we have selected papers that were deemed to be particularly relevant to two areas of research that figured prominently throughout the conference:

— corpus-based research into the grammar of Germanic, Slavic, and Romance languages
— methodological issues linked to corpus-based approaches to grammar and the application of corpus methods to related fields such as grammar education, the history of linguistics, and research on linguistic terminology.

These two focal points also shape the structure of the present volume, which is subdivided into two major parts:

— Part I: "Corpus-based Grammar Research"
— Part II: "Methodology and Application"

Each part contains a set of full-blown papers, which grew out of regular conference presentations, and a selection of shorter papers that correspond to poster presentations and present snapshots of current and ongoing research (grouped together under "Current Trends and Issues"). The thematic sections are introduced by the contributions of invited speakers at the conference: Anke Holler[6] and Alexandr Rosen, respectively. Part II contains a group of more application-oriented papers which starts with a chapter by another invited speaker, Susan Conrad. The volume ends with an epilogue by the final invited speaker at the conference, John Nerbonne. With the exception of the papers by the invited speakers, the longer as well as the shorter papers are ordered according to the languages of primary focus (with the sequence of Germanic – Romance – Slavic).

The subsequent overview of the content of the volume is divided according to the two areas of research mentioned above. We aimed at keeping the balance between these two areas throughout the volume, so that there is a due exchange between the description and analysis of specific languages/phenomena on the one hand, and methodological work and application-oriented approaches on the other hand. The papers are written in English or German as these were the conference languages. All contributions contain a short English abstract, which serves to indicate the theme of the paper in case the reader might not possess a profound knowledge of German (acknowledging the status of English as an academic lingua franca that most potential readers of this volume are familiar with).

Corpus-oriented Grammar Research

With the advent of large, annotated, searchable electronic corpora that can be accessed online, there has been a resurgence of interest in the use of corpus linguistic methods to study the grammar of natural languages.[7] As is well-known, corpus-based approaches to grammar are particularly useful in the study of linguistic variation. For the first time in the history of linguistics, researchers are able to draw on large amounts of data, which can be scrutinized by applying advanced statistical methods to discover even subtle fluctuations in the data.

6 In a chapter written together with Thomas Weskott.

7 It should perhaps be acknowledged that this general development has been foreshadowed by studies in historical linguistics, which have been assuming a pioneering role in corpus-based work on the grammar of natural languages, including the use of advanced statistical methods, cf. Pintzuk (2003) for an overview; more recent work includes e.g. Wallenberg (2009), Fruehwald et al. (2013), Ecay (2015), Kauhanen and Walkden (2017).

Moreover, this approach has proven to be very successful when it comes to the identification of factors (including both linguistic and extra-linguistic influencing parameters) that govern the distribution of variants in the corpus. This new, accessible, rich source of empirical evidence has also made available new possibilities to test and evaluate descriptive generalizations and the predictions of theoretical hypotheses, paving the way for more precise descriptions and better, more adequate theories. Both these points are amply demonstrated by the papers collected in this part of the volume.

However, the use of large corpora as empirical basis of grammar description and linguistic theory also raises a number of methodological and theoretical issues and challenges. In particular, we must be careful to avoid the potential fallacy of identifying the corpus with the grammatical system that we aim to describe. As large corpora consist of utterances produced by thousands, or even millions of speakers, they typically exhibit an amount of variation that is not found in any individual, including grammatical options that are incompatible with each other. Thus, a theoretical model that successfully captures the data in the corpus is not necessarily a valid description of an actual or even potential grammar in the mind of an individual speaker. To prevent wrong conclusions being drawn from the heterogeneous character of corpus data, a set of preparatory steps should be undertaken before we engage in the task of linguistic analysis (e.g. identification of phenomena and variants linked to extra-linguistic factors such as region, register etc.). In addition, certain questions arise concerning the nature of grammars constructed on the basis of corpus data. For example, one might ask whether relevant grammars represent an intersection or a union of the individual grammars that underlie the linguistic data collected in the corpus.

The contributions collected in this part of the volume all explore the use of corpus methods in the description and theoretical analysis of the grammar of natural languages, investigating a wide range of different phenomena in German, English, French, Spanish, Hungarian, and various Slavic languages. There is a set of recurring themes in the contributions on corpus-based research on grammar collected in this part of the volume:

— Language description and formal analyses should be based on a solid empirical foundation; moreover, corpora are a rich source for new and more precise empirical observations and descriptive generalizations. This is exemplified by basically all papers in this volume.
— Ideally, we should strive for a maximization of available evidence. That is, corpus data should be complemented by alternative methods (and vice versa), including experiments and introspection (cf. in particular the contributions by Holler and Weskott, Bader and Koukoulioti, and Elsner).

— Corpus-linguistic methods (together with the availability of parallel corpora) provide new options for comparative studies (cf. the contributions by Becker and Heck on the realization of aspect in various (Slavic) languages).
— Evidence from corpus studies can be used to evaluate and modify theoretical descriptions and models (cf. e.g. the papers by Holler and Weskott, Bader and Koukoulioti, Münzberg and Hansen-Morath, and Fricke and Tönnis).

The maximization of available evidence is a theme that repeatedly shows up in this collection. Ideally, linguists should not focus on a single empirical method, but rather should strive to seek converging evidence from a wide array of different data. This point is made very clearly in the contribution by Anke Holler and Thomas Weskott ("Implizite Verbkausalität im Korpus? – Eine Fallstudie"), who investigate the so-called implicit causality (IC) continuation bias, that is, the tendency to identify an anaphor with the stimulus argument rather than with the experiencer argument of a preceding verb. This effect is usually attributed to differences in salience between stimulus and experiencer arguments. By using the presence or absence of *von*-phrases ('by'-phrases) in passive clauses of German as another test case for measuring the relative salience of arguments, Holler and Weskott convincingly argue that experimental results should be complemented by, and checked against, evidence from actual language use collected in linguistic corpora. In this way, their contribution provides a link between corpus-based work on the grammar of languages and the methodological issues discussed in the second part of this volume.

In a similar vein, Markus Bader and Vasiliki Koukoulioti demonstrate in their paper "When Object-Subject Order is Preferred to Subject-Object Order: The Case of German Main and Relative Clauses" how corpus evidence can be used to shed light on issues pertaining to the conditions that govern the relative order of subject and direct object in main and relative clauses of German. They show that the corpus data corroborates earlier (experimental) findings, according to which orders where the object precedes the subject are the preferred option if the subject is a pronominal topic. Additionally, the possibility of OS-order is also influenced by properties of the object itself, namely its relation to the previous discourse and its categorical status (e.g., demonstrative vs. indefinite pronoun). The findings are then modelled making use of ranked violable constraints.

In their paper "*Die Wucht und Strömung war immens* – wie stark ist der Ellipseneffekt?" Franziska Münzberg and Sandra Hansen-Morath investigate agreement variation in connection with coordinated subjects in contemporary German. Focusing on singular noun phrases connected by *und* ('and'), they show that while plural agreement on the verb is the default choice, singular agreement becomes more likely when the determiner is elided in the second NP conjunct. In addition, they provide statistical evidence that the ellipsis effect is stronger

than other factors mentioned in the literature including subject individuation/agentivity.

The contribution by Tom Bossuyt, Ludovic de Cuypere, and Torsten Leuschner ("Emergence Phenomena in German *W-immer/auch*-Subordinators") is concerned with the distributional patterns of the German irrelevance particles *immer* ('ever') and *auch* ('also'), which in contrast to English *-ever* occur in multiple positions and combinations. Based on a sample of conditional and free relative clauses introduced by the *wh*-words *was* ('what') and *wer* ('who') (and their inflected forms), the paper offers a detailed description of the distribution of the particles (and combinations of them) and presents a functional analysis of the resulting patterns as a case of emergent grammar.

The paper by Jörg Didakowski and Nadja Radtke ("Deutsche Stützverbgefüge in Referenz- und Spezialkorpora: Vergleichsstudien mit dem DWDS-Wortprofil") deals with the distribution of light verb constructions (called "Stützverbgefüge" (SVG) by the authors) across different text types. The authors show how syntactic co-occurrences made available by the word profile of the Digital Dictionary of the German Language (Digitales Wörterbuch der deutschen Sprache, DWDS) can be used to identify potential SVGs. Subsequently, they present the results of three corpus studies that investigate the use of selected SVGs in different text types (newspapers, blogs, and a balanced corpus), focusing on the frequency, productivity, and diversity of SVGs. The results are then sorted by the density of predicate nouns, making use of three different association measures.

The paper by Oliver Wicher ("Corpus-Driven Lexical Grammar and the Aspect-Modality Interface: The Case of French Past Modal Constructions") investigates the interpretation of French past modal constructions such as *elle a pu rentrer* vs. *elle pouvait rentrer*, focusing on the so-called 'actuality entailment' effect: a perfect form of the root modal forces an interpretation where the event expressed by the complement takes place in the actual world. It is argued that the choice of different past tense forms is a matter of collostructional preference.

In the paper "Polar Verbless Clauses and Gapping Subordination in Spanish", Oscar Garcia-Marchena argues on the basis of empirical data taken from CORLE (Corpus of Contemporary Oral Spanish) that Spanish allows polar fragments and gapping in subordinate contexts, which are not permitted in English. More precisely, it is demonstrated that gapping, like other fragments, can only be embedded by verbal and non-verbal epistemic predicates, while polar verbless clauses are overall more frequent and can also be embedded by other types of predicates.

The contribution by Laura Becker ("Aspectuality in Hungarian, German, and Slavic. A Parallel Corpus Study") investigates whether Hungarian has a grammatical category of aspect, similar to e.g. the Slavic languages. Based on a parallel corpus of movie subtitles, verbal prefixation in Hungarian and German is

compared with the expression of aspect in Russian and Czech. It is shown that while Hungarian seems to pattern with Slavic languages for certain verb classes, aspectuality is largely determined by actionality in Hungarian, similar to German. From this, it is concluded that aspect is not a grammatical category in Hungarian.

The short paper by Daniela Elsner ("Empirisch basierte Überlegungen zu Ableitungen mit *-weise/-erweise*") combines corpus data with acceptability judgments to investigate adverbial word-formations with the formative *-(er)weise* in German. Based on the observation that formations with *-weise* differ from those with *-erweise* both in their interpretation and syntactic distribution, it is argued that the *-(er)weise* consists of two separate suffixes.

In "*Es ist dies* – A Special Use of German Prefield-*es*" Lea M. Fricke and Swantje Tönnis present a corpus study on a hitherto unstudied construction, where a prefield-*es* appears in combination with a demonstrative subject *dies* and a copula verb *ist*. It is shown that the construction is predominantly used in southern varieties of German. The authors then argue that the *Es ist dies* construction primarily serves to mark a topic shift and provide an analysis based on stochastic Optimality Theory (OT).

The short paper by Swantje Tönnis, Lea M. Fricke, and Alexander Schreiber ("Methodological Considerations on Testing Argument Asymmetry in German Cleft Sentences") investigates the relative frequency of subject and object *it*-clefts in German. By using a new method, the authors provide additional support for the claim that subject clefts are more frequent than object clefts in German. With its additional focus on methodological issues, the paper provides a link between the two major topics of this volume.

The short piece by Johanna Marie Poppek, Tibor Kiss, and Francis Jeffry Pelletier ("Kinds, Containers, Instances: Mass Nouns and Plurality") presents findings from a large-scale corpus study on the (surprisingly frequent) plural occurrences of mass nouns and so-called dual life nouns in English (which are both +count and +mass) and identifies a set of meaning shifts that result from pluralization that are linked to the countability class to which the noun belongs.

The contribution by Stefan Heck focuses on the category of aspect in Slavic ("A corpus study on verbal aspect in Czech, Polish and Russian imperatives"). Similar to Laura Becker, Heck assumes a comparative perspective, dealing with the realization of aspect in Czech, Polish, and Russian imperatives. It is shown that there are significant differences between Czech and Polish on the one side and Russian on the other.

In their contribution "Clitic Climbing and Stacked Infinitives in Bosnian, Croatian and Serbian – A Corpus-Driven Study", Björn Hansen, Zrinka Kolaković, and Edyta Jurkiewicz-Rohrbacher show that, in contrast to claims in the literature, clitic climbing is merely facultative in stacked infinitives of Bosnian,

Croatian and Serbian. In addition they identify a set of conditions that constrain the availability of clitic climbing in stacked infinitives.

Methodology and Application

The design and construction of corpora facilitating substantial linguistic research at different grammatical levels requires an intensive examination and reflection of a number of theoretical, technical, and practical issues, with corpus mark-up being one of the most crucial ones. In particular, linguistic annotation plays a decisive role in creating and exploring corpora by making linguistic information contained in the collected texts explicit and automatically accessible, the results of which make corpus studies reproducible and more accessible to others. Thereby, the steps and levels of linguistic annotation may incorporate various processes and linguistic phenomena related to phonological, morphosyntactic, semantic, or pragmatic aspects. While corpus annotation without doubt adds much value to a corpus, it always imposes one particular linguistic interpretation and is often inconsistent. Moreover, the quality of linguistic annotation may vary depending on whether it was performed manually, fully automatically, or semi-automatically. Certain types of corpora pose additional challenges and require a larger amount of manual work. The annotation of historical text collections usually calls for human philological expertise. Annotating corpora for the purposes of phonological analysis is particularly labor intensive. Moreover, the detection and annotation of phenomena such as phonemic contrasts and neutralization patterns, arguably requires that a lot of theoretical work be put into the annotation scheme, raising the question of whether potential benefits justify the effort. Different methodological issues related to the annotation of corpora, including dealing with historical texts, are addressed in the papers by Rosen, Raffelsiefen and Geumann, Tuggener and Businger, Bouma, Schauwecker and Stein, as well as Bilińska, Kwiecień, and Derwojedowa.

Over the past few decades, many interesting research methods have been developed within the analytical area. In particular, numerous statistical modeling techniques for language and speech have been extended, examined, and refined. Such techniques allow us not only to quantitatively describe, summarise, and systematise the features of our data collections (by the use of methods of descriptive statistics), but also to evaluate our data from the perspective of significance and, more importantly, to generalize (using statistical inference) from the properties observed in our datasets to the corresponding properties in the language as a whole. The majority of papers in this volume have integrated the application of basic or more sophisticated methods of descriptive or inferential statistics to corpus data into their analyses. The contribution by Tuggener and

Businger can serve as a perfect example where advanced statistical methods are used to unearth otherwise hidden patterns.

Corpora annotated for metadata and linguistic information have numerous applications. It is generally well known that they provide collections of examples for linguists (as demonstrated in Part I) and serve as data resources for lexicographers (cf. the Longman Dictionary of Contemporary English, the Duden dictionaries of the German language,[8] the Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts, DWDS[9]) and grammaticographers (cf. Biber et al. 1999, 2002; Huddleston and Pullum 2002, 2005). However, in this volume, we want to give a more comprehensive picture of the actual range of work carried out in the grammar and corpora setting, including lesser known and innovative areas of use. The application to disciplinary education and to foreign language teaching is addressed respectively in the papers by Conrad and Weber. The meta-grammatical use of corpora for automatic extraction of different kinds of information is demonstrated by Lang, Schneider, and Suchowolec with application to grammatical terminology, and by Busse, Gather, and Kleiber for information relevant to the history of science.[10]

The contribution by Alexandr Rosen ("Coping with Unruly Language: Non-Standard Usage in a Corpus") is concerned with non-canonical linguistic expressions, which exhibit irregular (non-compositional) semantics, syntax, morphology, pragmatics, and/or phonology and may involve phenomena such as performance errors, creative coinages, or emerging appearances (multi-word expressions are a perfect example). Due to the fact that non-standard language does not obey general grammar rules, it cannot be handled using categories, methods, and tools developed for canonical language. Rosen suggests two ways to approach this problem: the first approach applies to the design of an annotation scheme for Czech learner corpora, and the second one to the grammar-checked annotation of a parsebank.

The paper by Renate Raffelsiefen and Anja Geumann ("Phonological Analysis at the Word Level: The Role of Corpora") addresses the question to what extent a corpus-driven approach can yield insights into phonemic structures and phonological systems. Focusing on quality and quantity contrasts in the vowel system of German, the authors draw on evidence from various sources and phenomena, including acronyms, loanwords, and speech errors to argue for a more

8 E.g. Duden (2017) or Duden online.

9 <https://www.dwds.de/> (7.5.2018).

10 For sake of completeness, it should be added that linguistic corpora are also extensively used for training different NLP tools, such as speech recognizers, statistical part-of-speech taggers, and parsers, as well as example-based and statistical machine translation systems.

theory-driven constraint-based approach to phonology. In addition, they discuss how different corpus resources can be used as an empirical basis for phonological analysis.

The paper by Don Tuggener and Martin Businger ("Needles in Haystacks: Semi-Automatic Identification of Regional Grammatical Variation in Standard German") presents a semi-automatic method to identify regional variation in the grammar of Standard German in the domains of inflection, word formation and valency. It is demonstrated that the proposed method not only allows us to identify a known variation, but also makes it possible to discover language variants that have not yet been attested.

The paper by Gosse Bouma ("Corpus-Evidence for True Long-Distance Dependencies in Dutch") discusses problems of finding corpus evidence for long-distance dependency phenomena, which is a well-known challenge for statistical parsers. It presents relevant results from an automatically annotated treebank for Dutch (Lassy Large) and argues that this corpus is sufficiently large and heterogeneous to serve as an adequate data source for non-local phenomena. The results of the corpus queries suggest that in Dutch, true long-distance dependencies are rare and have limited productivity; additionally, they seem to involve collocational effects.

The problem of automatic grammatical annotation of non-standardised languages is the topic of the contribution by Yela Schauwecker and Achim Stein ("Automatic Morphosyntactic and Dependency Annotation of the Anglo-Norman Text Database"). The paper discusses the annotation of the Anglo-Norman text database, addressing a number of linguistic and extra-linguistic peculiarities related to this specific type of historical data. They show how the data from Anglo-Norman (a variety of Old French) can be normalised and how a dependency parser developed for Old French can then be applied to the normalised Anglo-Norman data.

Related problems pertaining to the automatic annotation of grammatical properties in historical texts are dealt with in the contribution by Joanna Bilińska, Monika Kwiecień, and Magdalena Derwojedowa ("Microcorpus of Nineteenth-Century Polish"). The paper shows how a morphological analyser developed for contemporary Polish can be adapted to process historical inflection and spelling in a small corpus of nineteenth-century Polish texts.

The use of corpus linguistic methods in the field of applied linguistics is showcased by Susan Conrad's contribution "Beyond Grammar Description: Applying Corpus Analysis to Disciplinary Education", in which she describes an interdisciplinary project concerning civil engineering writing. Starting from corpus-based grammar-related analyses of student and practitioner writing, specific teaching materials are developed to improve the writing skills of engineering students. Additional corpus analyses are used to evaluate the impact of the materials on student writing.

In the application-oriented short paper "Grammatik und Lernerkorpora: Eine korpusorientierte Untersuchung von Präpositionalphrasen im deutschen MERLIN-Korpus", Tassja Weber's analysis of the German learner corpus MERLIN shows that learners have greater problems with prepositional objects (PO), where the preposition has only weak semantic content, than with adverbial PPs, where the preposition has a more specific meaning, as learners more often erroneously omit the preposition in POs.

In their short paper "Extracting Specialized Terminology from Linguistic Corpora", Christian Lang, Roman Schneider, and Karolina Suchowolec compare different methods for extracting German grammatical terminology, demonstrating the importance of unigrams in grammar writing. They show that corpus comparing methods outperform alternative methods.

The pilot study by Beatrix Busse, Kirsten Gather, and Ingo Kleiber "Assessing the Connections between English Grammarians of the Nineteenth Century – A Corpus-Based Network Analysis" investigates a corpus of nineteenth-century English grammars, focusing on the transition from prescriptive to descriptive grammar writing. The paper shows that this paradigmatic change can be traced both in the network of grammarians' references and in the way terms like *prescriptive* and *descriptive* are used in the grammars.

In its condensed brevity, the above overview highlights the fact that the present collection covers a wide array of different languages, topics, and methodological approaches. This can, hopefully, indicate the vast spectrum of the productive research work in the grammar and corpora setting. With any luck, the volume will help to spread relevant insights across the boundaries of individual disciplines, philologies, and theoretical frameworks, and in this way further an interdisciplinary and collaborative approach to the investigation of language. It reveals, in any case, that corpus linguistic methods are already entrenched and technically advanced in the grammar research of languages focused on in this book. Today, corpora are built, edited, annotated, searched, and analysed with the aid of a computer and are so commonly available that grammar research without corpus linguistic methods has become almost unthinkable. Consequently, in the future, there will be less need to promote corpus linguistic methods in grammar research, and one can think of shifting the profile of the next *Grammar and Corpora* conferences from monitoring how corpus linguistic methods trigger new insights in very different areas of grammar, to focusing on selected methodical issues and/or specific subfields of grammar. Finally, after having read all the manifold contributions about grammar and corpora, a lot of metalinguistic questions might arise in the reader's mind, e.g. about the theoretical status of corpus research on grammar, about its interdisciplinary position, or about its genesis and future development. At least some of these questions will be seized

on in the epilogue of the book, where John Nerbonne comprehensively reflects on the interplay of grammatical theory, corpus linguistics, and computational linguistics that has been conditioning the corpus approach to grammar in the last decades.

At this point, we would like to use the opportunity to direct some words of sincere gratitude and appreciation to several people without whom this volume could not have been accomplished. First of all, due words of thanks go to the authors for their contributions and for their meeting tight publication deadlines and to all the members of the advisory board for active help. We are also very grateful to the staff of Heidelberg University Publishing, who supported us extremely competently in all editorial matters and offered us the opportunity to publish the volume in multiple formats.

References

Biber, Douglas, Susan Conrad and Geoffrey Leech. 2002. *Longman Student Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Digitales Wörterbuch der deutschen Sprache des 20. Jahrhunderts (DWDS). https://www.dwds.de/ (3.11.2017).

Duden. 2017. *Duden: Die deutsche Rechtschreibung*. 27th edn. Berlin: Dudenverlag.

Duden online http://www.duden.de/ (3.11.2017).

Ecay, Aaron. 2015. A multi-step analysis of the evolution of English *do*-support. Ph.D. dissertation, University of Pennsylvania, Philadelphia.

Fruehwald, Josef, Jonathan Gress-Wright and Joel Wallenberg. 2013. Phonological rule change: The Constant Rate Effect. In Seda Kan, Claire Moore-Cantwell and Robert Staubs (eds.), *NELS 40. Proceedings of the 40th Annual Meeting of the North East Linguistic Society* (vol. 1), 219–230. Amherst, MA: University of Massachusetts, GLSA Publications.

Hansen-Morath, Sandra and Sascha Wolfer. 2017. Standardisierte statistische Auswertung von Korpusdaten im Projekt *Korpusgrammatik* (KoGra-R). In Konopka, Marek and Angelika Wöllstein (eds.), *Grammatische Variation. Empirische Zugänge und theoretische Modellierung*. (= Jahrbuch des Instituts für Deutsche Sprache 2016), 345–356. Berlin/Boston: de Gruyter.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: CUP.

Huddleston, Rodney and Geoffrey K. Pullum. 2005. *A Student's Introduction to English Grammar*. Cambridge: CUP.

Kauhanen, Henri and George Walkden. 2017. Deriving the Constant Rate Effect. *Natural Language and Linguistic Theory* <https://doi.org/10.1007/s11049-017-9380-1>.

Konopka, Marek, Jacqueline Kubczak, Christian Mair, František Štícha and Ulrich H. Waßner (eds.). 2011. *Grammatik und Korpora 2009. Dritte Internationale Konferenz. Grammar and Corpora 2009. Third International Conference* (22.–24.09.2009, Mannheim). Tübingen: Narr.

Kupietz, Marc and Harald Lüngen. 2014. Recent Developments in DeReKo. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2378–2385. Reykjavik: ELRA.

Münzberg, Franziska. 2016. Grammar and Corpora 2016 – Korpuslinguistinnen und -linguisten zu Gast in Mannheim. *Sprachreport* 33(1), 40–42.

Pintzuk. Susan. 2003. Varationist approaches to syntactic change. In Brian D. Joseph and Richard D. Janda (eds.), *The Handbook of Historical Linguistics,* 509–528. Oxford: Blackwell.

Przepiórkowski, Adam, Zygmunt Krynicki, Łukasz Dębowski, Marcin Woliński, Daniel Janus and Piotr Bański. 2004. A search tool for corpora with positional tagsets and ambiguities. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)* <http://www.lrec-conf.org/proceedings/lrec2004/pdf/275.pdf>.

Rosenfeld, Viktor. 2010. An implementation of the Annis 2 query language. Technical report, Humboldt-Universität zu Berlin.

Schäfer, Roland. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of Challenges in the Management of Large Corpora (CMLC-3)* (IDS publication server), 28–34 <https://ids-pub.bsz-bw.de/files/3826/Schaefer_Processing_and_querying_large_web_corpora_2015.pdf>.

Schäfer, Roland. 2016. CommonCOW: massively huge web corpora from CommonCrawl data and a method to distribute them freely under restrictive EU copyright laws. In *Proceedings of LREC 2016* <http://www.lrec-conf.org/proceedings/lrec2016/pdf/960_Paper.pdf>.

Štícha, František and Mirjam Fried (eds.). 2008. *Grammar and Corpora 2007. Grammatik a Korpus 2007* (25.–27.09.2007, Liblice). Prague: Academia.

Štícha, František and Josef Šimandl (eds.). 2007. *Grammatika a korpus 2005. Grammar and Corpora 2005* (23.–25.11.2005, Prague). Prague: The Institute of the Czech Language of the Academy of Sciences of the Czech Republic.

Wallenberg, Joel. 2009. Antisymmetry and the Conservation of C-Command: scrambling and phrase structure in synchronic and diachronic perspective. Doctoral Dissertation, University of Pennsylvania.

I. Corpus-Based Grammar Research

Anke Holler, Thomas Weskott

Implizite Verbkausalität im Korpus? – Eine Fallstudie

Abstract Experimental psycholinguists have studied the so-called implicit causality bias for more than forty years and have mostly attributed it to an effect of argument structure of a certain class of interpersonal verbs on subsequent anaphor resolution; most accounts attribute this effect to differences in salience between the arguments of the respective verbs. This article reports a corpus study on passive sentences for two classes of implicit causality verbs and puts the salience hypothesis to test. By tracing the implicit causality bias in corpora and taking into consideration a wider variety of contexts than usually employed in experiments, we want to scrutinize the ecological validity of the experimental results. From a more general point of view, the aim of the article is to exemplify how results from different methodological approaches, i.e. experiments and corpus search, can be brought to bear on our understanding of a grammatical phenomenon.

Keywords Implicit causality bias, psych verbs, ecological validity, passives

1 Einleitung

Die psycholinguistische Forschung widmet sich seit mehr als vier Jahrzehnten den beobachtbaren Effekten der sogenannten impliziten Verbkausalität. Dieses Konzept erfasst eine inhärent lexikalische Eigenschaft von transitiven Verben, die eines der beiden Argumente des Verbs (üblicherweise das Agens) als Verursacher des ausgedrückten Sachverhalts ausweist. Damit einher geht eine erhöhte Salienz dieses Arguments, wie zahlreiche experimentelle Studien gezeigt haben. Üblicherweise wird dabei mithilfe von Satzvervollständigungsaufgaben der Form [Argument$_1$ Verb Argument$_2$, *weil* Pronomen …] ermittelt, welches Argument präferiert pronominal wiederaufgenommen wird. Die Präferenz bezüglich des anaphorischen Bezugs wird auch als *implicit causality bias* (kurz: IC-Bias) bezeichnet und in der Regel auf Asymmetrien in der Argumentstruktur der beteiligten Verben zurückgeführt. Diese Klasse wird in der psycholinguistischen

Literatur üblicherweise als interpersonale Verben oder IC-Verben bezeichnet. Das Hauptaugenmerk liegt aber genau genommen auf einer Teilmenge dieser Verben, die mit der linguistischen Klasse der psychischen Verben nahezu deckungsgleich ist. Obwohl dem IC-Bias eine prominente Rolle in der psycholinguistischen Literatur zukommt, ist die Frage bisher nicht thematisiert worden, ob und inwieweit die beobachteten Effekte des IC-Bias ökologisch valide sind, d. h. außerhalb eines sorgfältig kontrollierten Experiments Bestand haben. Da die Datenerhebung zudem bisher nur in Experimentalstudien mit einer sehr eingeschränkten Menge von Stimuli erfolgt ist, wurden mögliche grammatische Einflüsse auf den IC-Bias kaum berücksichtigt. Nennenswerte Ausnahmen stellen die Arbeiten von Corrigan (1988) zur Belebtheit oder von Brown & Fish (1983) zur Definitheit dar.

Korpora haben den Vorzug, dass sie gebrauchsbasiert sind und daher für eine sprachliche Zielgröße u. a. Auskunft über ihr Vorkommen, die Häufigkeit ihres Vorkommens und/oder die sprachliche Umgebung ihres Vorkommens geben können. Unter der Annahme, dass Korpusdaten somit einen direkteren Zugriff auf das multivariate Zusammenspiel grammatischer Eigenschaften erlauben als experimentelle Befunde, wollen wir untersuchen, ob sich der IC-Bias in Korpora abbilden lässt. Dabei wollen wir anhand einer Fallstudie zur impliziten Verbkausalität in passivierten Sätzen zeigen, dass Korpusdaten zur Überprüfung der ökologischen Validität von Experimentaldaten herangezogen werden können. Mit anderen Worten: Wir möchten der Frage nachgehen, inwieweit Korpusdaten geeignet sind, aufzuklären, ob sich bestimmte sprachliche Entitäten im Gebrauch genauso verhalten wie im wohlkontrollierten Experiment. Wir nehmen an, dass nur dann, wenn die Antwort auf diese Frage positiv ausfällt, es gerechtfertigt ist, die Ergebnisse einer experimentellen linguistischen Untersuchung als ökologisch valide einzustufen. Unser Vorschlag ist also, Korpusfrequenzdaten zu nutzen, um für experimentelle Settings in der Linguistik den Grad ihrer Approximation an wirkliche sprachliche Gegebenheiten zu bestimmen. Der Aufsatz schließt damit an die grundsätzliche Diskussion darüber an, wie Ergebnisse, die durch verschiedene methodische Zugänge zustande kommen, zusammengenommen für die Analyse eines grammatischen Phänomens geltend gemacht werden können. In unserem konkreten Fall heißt dies, die bestimmten psychischen Verben eigene Salienzmarkierung ihrer Argumente, die experimentell im IC-Bias ihren Ausdruck findet, auch durch korpusbasierte Gebrauchsdaten zu belegen.

Mit unserer Studie verfolgen wir ausdrücklich nicht das Ziel, die linguistische Theoriebildung fortzuentwickeln und beispielsweise zu grundlegenden Fragen hinsichtlich der Motivation für Passivierung oder zur sprachtheoretischen Behandlung von psychischen Verben beizutragen, sondern es geht uns vorrangig um ein methodologisches Problem. Indem wir exemplarisch aufzeigen, wie mittels eines im Korpus leicht zugänglichen linguistischen Ausdruckstyps (i. e. der

Passivierung) ein psycholinguistischer Befund (i. e. der IC-Bias) validiert werden kann, wollen wir einen Beitrag zur Beantwortung der generellen Frage leisten, in welcher Weise Korpusdaten und Experimentaldaten einander sinnvoll komplementieren können.

Der Aufsatz gliedert sich wie folgt: Nach einer ausführlichen Beschreibung des IC-Bias, seiner psycholinguistischen Fundierung und der sich daraus ergebenden Forschungsfrage bezüglich der ökologischen Validität der Experimentaldaten im nachfolgenden Abschnitt 2 werden wir in Abschnitt 3 die Korpusstudie vorstellen, die wir durchgeführt haben, um die in der Psycholinguistik zur Erklärung des IC-Bias gängige Salienzhypothese anhand von Korpusfrequenzdaten zu überprüfen. Im abschließenden Abschnitt 4 werden wir die Ergebnisse dieser Studie problematisieren und zur Diskussion über die ökologische Validität der psycholinguistischen Befunde in Beziehung setzen.

2 Fragestellung

Im Zuge der empirischen Wende in der Sprachwissenschaft haben experimentell erhobene sprachliche Daten für die linguistische Theoriebildung an Bedeutung gewonnen. Verglichen mit anderen linguistischen Datentypen zeichnen sie sich vor allem dadurch aus, dass sie aus kontrolliert durchgeführten Experimenten stammen, in denen sprachliche Phänomene hypothesengeleitet hinsichtlich einer oder mehrerer zuvor festgelegter Kriterien nach einer mehr oder minder verbindlich vorgegebenen Prozedur untersucht werden. Auf diese Weise ist im Idealfall weitestgehend sichergestellt, dass die bezüglich des jeweiligen sprachlichen Phänomens beobachteten Effekte tatsächlich auf die untersuchten Faktoren bezogen werden können und nicht etwa Resultat anderer, zuvor nicht berücksichtigter Einflussgrößen sind. Gleichzeitig stellt sich aber auch die Frage, inwieweit die so gewonnenen Daten überhaupt die realen Gegebenheiten repräsentieren. Eine weitergehende Frage ist dann, ob experimentelle Bedingungen die Gegebenheiten des natürlichen Sprachgebrauchs approximieren können und wollen. Demgegenüber liefern Korpora Evidenz, die nicht aus einer kontrollierten experimentellen Situation stammt und damit nicht aus einer Vorauswahl der zu betrachtenden Eigenschaften resultiert; vielmehr kann in Korpusstudien eine Zufallsauswahl an Textbelegen getroffen werden, ohne im Vorhinein die Anzahl der Eigenschaften festlegen oder auch nur kennen zu müssen. Insofern ist naheliegend zu fragen, ob beide Datentypen, Experimentaldaten und Korpusdaten, sinnvoll miteinander kombiniert werden können, sodass insgesamt in Bezug auf ein sprachliches Phänomen ein vollständigeres und verlässlicheres Bild entsteht. Dieses Zusammenspiel von experimentell beobachteten Effekten einerseits und korpusbasiert erhobenen Befunden andererseits wollen wir anhand des IC-Bias

einer Subklasse der Psychverben eingehender betrachten. In einem ersten Schritt werden wir dazu nachfolgend einige methodologische Aspekte, insbesondere hinsichtlich der ökologischen Validität und des IC-Bias, diskutieren und danach die Salienzhypothese für den IC-Bias, die als Ausgangspunkt für die in Abschnitt 3 dargestellte Korpusstudie dient, einführen sowie die von uns verwendete Operationalisierung mittels der Passivkonstruktion motivieren.

2.1 Experiment vs. Korpus

Neben explorierenden, hypothesengenerierenden Experimenten sind es vor allem hypothesentestende Experimente, denen in der Linguistik und Psycholinguistik eine gewichtige Rolle zukommt. Bei diesem Experimenttyp soll anhand einer Stichprobe von Sprechern und von (morphologischen, syntaktischen, semantischen) Instanzen eines Ausdruckstyps (den *Items*) eine Hypothese geprüft werden oder, genauer gesagt, es soll die zugehörige Nullhypothese anhand der Daten der Stichprobe verworfen und damit die empirische Hypothese angenommen werden. Der entscheidende Schritt ist dabei die Inferenz vom Nicht-Zutreffen der Nullhypothese (zum Beispiel: Es besteht hinsichtlich der Mittelwerte einer abhängigen Variablen v kein Unterschied zwischen den beiden Ausprägungen eines Faktors A, a_1 und a_2, d. h. $\bar{x}_{v(a1)} = \bar{x}_{v(a2)}$.) in der Stichprobe auf das Zutreffen der empirischen Hypothese (zum Beispiel: Es besteht ein Unterschied zwischen den Mittelwerten, $\bar{\mu}_{v(a1)} \neq \bar{\mu}_{v(a2)}$ in der Population.), bei dem die Irrtumswahrscheinlichkeit üblicherweise auf 5 % festgelegt ist. Die Kombination von systematischer Manipulation eines Faktors bei gleichzeitiger Minimierung des Einflusses von Störvariablen durch Kontrolle über die experimentelle Situation mit inferenzstatistischen Verfahren stellt sicher, dass der Unterschied zwischen den erhobenen Mittelwerten der abhängigen Variablen mit einer Irrtumswahrscheinlichkeit von $p < .05$ tatsächlich auf den Faktor zurückzuführen ist, genauer gesagt, dass die Wahrscheinlichkeit, dass das Unterschiedsmuster beobachtbar ist, wenn in der Population die Nullhypothese gilt, kleiner 5 % ist. Bei den meisten inferenzstatistischen Verfahren handelt es sich dabei um sogenannte *parametrische* Verfahren, d. h. es werden anhand von Stichprobenparametern (wie beispielsweise dem Mittelwertunterschied und dessen Streuung beim *t*-Test) bestimmte in der Population geltende Parameter (der in der Population geltende Mittelwertunterschied sowie dessen Streuung) geschätzt.

Ein Nachteil dieser Kombination von kontrolliertem Experiment und hypothesenprüfender Inferenzstatistik ist, dass der Anzahl der Faktoren, die in einem Experiment manipuliert werden können, durch die Anforderungen der statistischen Verfahren und deren Auswirkung auf Itemanzahl und damit letztlich auf die Belastbarkeit der Probanden recht enge Grenzen gesetzt sind: Um

eine verlässliche Schätzung der Populationsparameter für eine experimentelle Bedingung zu erhalten (wie beispielsweise eines Ratingmittelwertes und seiner Streuung über Probanden bzw. Items hinweg), braucht man für diese Bedingung mindestens 6 Beobachtungen (d. h. 6 Probanden / Items pro Bedingung). Für ein Experiment, das zwei zweistufige Faktoren miteinander kreuzt, ergibt sich also nach dieser Faustregel schon die Notwendigkeit, 24 Probanden und 24 Items zu testen; zusammen mit einer weiteren „Goldenen Regel" des Experimentierens, und zwar, dass das Verhältnis von Filleritems (d. h. Ablenkern von der experimentellen Fragestellung) zu experimentellen Items 2:1 sein sollte, ergäben sich daraus schon 24 + 48 = 72 Items, die von 24 Probanden bearbeitet werden müssen. Jeder weitere zweistufige Faktor, der hinzugenommen wird, verdoppelt diese Anforderungen; spätestens bei einem Experiment, das die gegenseitigen Abhängigkeiten von vier zweistufigen Faktoren testet (also, in der Redeweise der Experimentalpsychologie, bei einem 2 × 2 × 2 × 2-Design), erschöpfen sich – im wahrsten Sinne des Wortes – die Möglichkeiten der experimentellen Überprüfung in *einem* Experiment. Die Anzahl der zu bearbeitenden Items übersteigt dann (bei ca. 300 Items) bereits die Menge, die Probanden üblicherweise zugemutet werden kann, ohne dass die Datenqualität durch Ermüdungserscheinungen leidet.

Demgegenüber haben Korpusstudien den Vorteil, dass sich satzbasiert große Mengen von Daten hinsichtlich einer potenziell beliebigen Anzahl von Faktoren annotieren lassen und es damit durchaus möglich ist, um beim Beispiel zu bleiben, gleich große Substichproben für die 16 Zellen eines 2 × 2 × 2 × 2-Designs zu erhalten. Anhand einer solchen Stichprobe für ein gegebenes Phänomen ließe sich dann, ganz analog zum inferenzstatistischen Verfahren beim Experimentieren, von einem Befundmuster für die Stichprobe (die annotierten Korpusbelege) auf die Population (das Korpus als Ganzes oder gar das Genre, das das Korpus repräsentiert) schlussfolgern. Allerdings wäre hier eine Anzahl von n=6 Beobachtungen pro Ausprägungen eines Annotationsmerkmals schon deshalb nicht hinreichend, weil in Korpusbelegen – mangels Kontrolle über die Faktoren – außer den annotierten Eigenschaften immer noch weitere Eigenschaften zwischen den Items variieren; um eine systematische Konfundierung mit diesen Eigenschaften zu vermeiden, muss die Stichprobengröße entsprechend angepasst werden, was im Falle seltener Phänomene (d. h. lexikalisch und/oder syntaktisch restringierter Vorkommen der Form) problematisch sein kann.

Zusammenfassend lässt sich konstatieren, dass experimentelle Methoden aufgrund der systematischen Kontrolle von Faktoren Generalisierbarkeit um den Preis der Natürlichkeit des jeweiligen experimentellen Settings erkaufen, während korpusanalytische Methoden weitestgehend natürliche Daten liefern, allerdings um den Preis der systematischen Kontrolle der Einflussgrößen.

2.2 Ökologische Validität

Jede für ein Experiment gewählte Verfahrensweise hat letztlich ihre Grenzen im Einfluss der jeweiligen Versuchspersonen und der situativen Umgebung. Auch durch eine möglichst wirklichkeitsnahe Gestaltung eines experimentellen Settings können diese Einflussgrößen nur bis zu einem bestimmten Grad minimiert werden. Insofern stellt sich für jedes experimentell gewonnene Ergebnis die Frage nach seiner Generalisierbarkeit, die letztlich auch davon abhängt, inwieweit die Experimentalanordnung mit den natürlichen Gegebenheiten übereinstimmt. Dieser Zusammenhang ist in der Psychologie mit dem Konzept der ökologischen Repräsentativität bzw. ökologischen Validität erfasst worden, die auch als „die empirische Gültigkeit einer psychologischen Aussage für das Alltagsgeschehen" (Dorsch 2013) beschrieben wird. Es geht also um das Verhältnis zwischen der konstruierten und in diesem Sinne artifiziellen Datenerhebung im Labor und den Gegebenheiten in einer natürlichen Lebensumgebung. Bezogen auf sprachliche Daten ist also zu fragen, worin die für die Gültigkeit der Daten relevanten natürlichen Lebensumstände bestehen und wie ihre Bedingungen bestimmt werden können. Auf den ersten Blick ist es naheliegend, den alltäglichen Sprachgebrauch als das für die ökologische Validität relevante „Biotop" (Pawlik 1976) anzusehen. Doch um das Ausmaß der Entsprechung und damit die Vorhersageleistung der kontrollierten Datenerhebungen in der jeweiligen Kriteriensituation ermitteln zu können, bedarf es einer Methode, die es ermöglicht, die „unverstellten" natürlichen Sprachdaten als Vergleichsgrundlage zu archivieren. Dies kann durch die Erstellung und Aufbereitung von großen Korpora geleistet werden. (Eine Analogie besteht hier möglicherweise zu der Unterscheidung zwischen nicht-intervenierenden/beobachtenden vs. experimentellen Verfahren in der Psychologie.) Da Korpusdaten, die üblicherweise auch als Produktionsdaten beurteilt werden, aus sprachlichem Handeln in einer konkreten Situation resultieren, schlagen wir vor, Korpora auszuwerten, um die empirische Repräsentativität von experimentell gewonnenen Sprachdaten zu überprüfen. Allerdings können Korpora den sprachlichen Alltag nur approximieren, da Korpusdaten im Zuge der Auswahl und Aufbereitung bereits kategorisiert und interpretiert werden. Hinzu kommt, dass in Korpora die Textsorte „Zeitungstext" dominiert und die dort enthaltenen Äußerungen in den meisten Fällen editiert sind. Ein vollkommen unverstellter, natürlicher Zugang zu Sprachdaten, wie es das Desiderat der ökologischen Validität erfordert, ist damit auch durch Korpora nicht vollständig gegeben. (Um die Analogie zur Psychologie aufzugreifen: Dies ist mutatis mutandis vergleichbar mit dem Beobachterparadoxon in der (Sozial-) Psychologie.) Unter der Annahme, dass Korpusdaten einen direkteren Zugriff auf grammatische Gegebenheiten erlauben als experimentelle Befunde, scheinen Korpora aber gut geeignet, um die Übertragbarkeit von Laborergebnissen

auf tatsächliche sprachliche Äußerungssituationen festzustellen und damit die ökologische Validität beobachteter Effekte zu bestimmen. Inwieweit dieser Zusammenhang hergestellt werden kann, wollen wir in einer Fallstudie ergründen. Anhand des IC-Bias interpersonaler Verben, eines in der Psycholinguistik fest etablierten Befundes, werden wir untersuchen, inwieweit er sich in Korpora abbilden lässt und welchen Rückschluss dies auf seine ökologische Validität zulässt.

2.3 IC-Bias der psychischen Verben

Unter dem Etikett *implizite Kausalität* wird in der experimentellen Psychologie und der Psycholinguistik das Phänomen verhandelt, dass die Information, die der Zuschreibung von Ursache und Wirkung in Ereignissen zugrunde liegt (die kausale Attribution zu Partizipanten eines Ereignisses), in den lexikalischen Einträgen von Verben implizit enkodiert ist. Für agentiv-kausative *accomplishment-* oder *achievement*-Verben wie beispielsweise *schlagen* lässt sich diese Zuschreibung ohne Weiteres aus der syntaktischen Funktion der Argumente ablesen: Der Referent des Subjekts des SCHLAGEN-Ereignisses ist – qua Agens – ursächlich für die Wirkung, welche im Affiziertsein des Referenten des direkten Objekts durch das SCHLAGEN-Ereignis (als Bestandteil des Resultatszustandes) besteht. Weniger klar und nicht aus der syntaktischen Funktion der Argumente ableitbar ist die Zuschreibung dieser kausalen Rollen[1] für psychische Verben. Den paradigmatischen Fall bildet das Paar *fürchten* vs. *ängstigen*:

(1) a. Siggi ängstigt Erwin.
    b. Erwin fürchtet Siggi.

In einer Situation, in der Siggi Erwin ängstigt, ist es wahrscheinlich gleichzeitig der Fall, dass Erwin Siggi fürchtet. Hieraus ergibt sich schon, dass die kausale Rolle des Verursachers nicht am Subjekt festgemacht werden kann; vielmehr scheint die Rolle des Verursachers in (1.a) beim Referenten des Subjekts, in (1.b) aber bei dem des direkten Objekts verortbar zu sein (also beide Male bei Siggi); die Wirkung, also der psychische Zustand des Fürchtens bzw. Geängstigtseins, ist in (1.a) im Objekt-, in (1.b) aber im Subjektreferenten zu finden (also beide Male bei Erwin). Es liegt nahe, die kausalen Rollen an die jeweiligen thematischen Rollen zu knüpfen: der die Furcht/Angst erfahrende

1 Die Redeweise von Kausalität ist hier selbstverständlich nicht im Sinne naturgesetzlicher Determiniertheit der Wirkung durch die Ursache zu verstehen.

oder erlebende Erwin ist in beiden Fällen eher Teil der Wirkung, während der Furcht/Angst einflößende Siggi eher Teil der Ursache ist. Dies mag Garvey & Caramazza (1974), die diese Beobachtung erstmals experimentell absicherten, dazu bewogen haben, von *causal valence* zu sprechen. In neuerer Terminologie ist es bei psychischen Verben die Rolle des Stimulus, der man kausale Wirkmächtigkeit, und die des Experiencers, der man die (psychische) Affiziertheit durch das Verbereignis zusprechen würde (siehe Postal, 1971; Belletti & Rizzi, 1988 und die darauffolgende Literatur). Wir schließen uns dieser Nomenklatur an und werden, wie das in der psycholinguistischen Literatur zum Thema üblich ist, psychische Verben wie ä*ngstigen*, *nerven*, *überraschen* im Folgenden als Stimulus-Experiencer-Verben (kurz: SE-Verben) bezeichnen, und psychische Verben wie *fürchten*, *hassen*, *bemerken* als Experiencer-Stimulus-Verben (kurz: ES-Verben).[2]

Garvey & Caramazza (1974) kommt das Verdienst zu, auf die Asymmetrie hinsichtlich der kausalen Valenz aufmerksam gemacht zu haben; ihr Squib in *Linguistic Inquiry* kann als Anfangspunkt eines regelrechten Industriezweigs der Psycholinguistik und der experimentellen Sprach- und Sozialpsychologie gesehen werden. Die Asymmetrie der kausalen Valenz bei psychischen Verben ist in den letzten etwa vierzig Jahren für unzählige Sprachen und Sprechergruppen (Sprachlerner, Erwachsene) nachgewiesen worden, und zwar mithilfe verschiedenster experimenteller Paradigmen sowohl psycholinguistischer wie auch kognitions- und sozialpsychologischer Provenienz: vom einfachen *Rating* der kausalen Wirkmächtigkeit für die beiden Referenten über Attributions- und Satzvervollständigungsaufgaben bis zu komplexen Online-Erhebungsmethoden wie *Eyetracking* beim Lesen und im *Visual World*-Paradigma (siehe Pickering & Majid 2007 und Hartshorne 2013 für einen Überblick). Dabei erwies sich vor allem ein Effekt als äußerst robust, der aus (psycho-)linguistischer Sicht interessant ist: Präsentiert man Probanden hinsichtlich eines Pronomens wie *er* ambige Satzfragmente der Form in (2) und bittet sie, diese zu vervollständigen, so zeigt sich, dass in den Satzvervollständigungen das ambige Pronomen *er* koreferent mit dem Stimulus-Argument (*Siggi*) ist.

(2) a. Siggi ängstigte Erwin, weil er …
… schon wieder die Wildschweinmaske trug.
(*er* = $\text{Siggi}_{\text{Stimulus}}$)

2 Es sei der Leserin überlassen, diese durch die in der syntaktischen Literatur gängige Terminologie zu ersetzen, derzufolge *ängstigen* ein Object-Experiencer-Verb, *fürchten* aber ein Subject-Experiencer-Verb ist.

b. Erwin fürchtete Siggi, weil er …
… schon wieder die Wildschweinmaske trug.
(*er* = $\text{Siggi}_{\text{Stimulus}}$)

Diese Präferenz zugunsten des Stimulus-Arguments bei der Anaphernresolution ist als ein empirisches Korrelat des IC-Bias in die Literatur eingegangen. Naturgemäß war es eher der psycholinguistische als der sozialpsychologische Forschungszweig, der sich diesem Befund gewidmet hat, und zwar nicht zuletzt deshalb, weil Fälle wie (2.b) offensichtlich den gängigen Heuristiken zuwiderlaufen, die im Sprachverstehen wie in der Sprachproduktion für die Anaphernresolution angesetzt werden: Subjektpräferenz und First-Mention-Strategie, denen zufolge das Antezedens von *er* in (2.b) *Erwin* sein müsste (qua Subjekt bzw. ersterwähntem Referenten).

Der IC-Bias zugunsten des Stimulus-Arguments bei der Anaphernresolution beläuft sich den Metaanalysen von Ferstl, Garnham & Manoulidou (2011) und Hartshorne (2013) zufolge auf ca. 85 % (mit Schwankungen in Abhängigkeit vom betrachteten Verb bzw. Verbpaar). In der psycholinguistischen Literatur besteht weitestgehend Einigkeit darüber, dass es die Asymmetrie auf der Ebene der Argumentstruktur ist, die für den IC-Bias verantwortlich ist; siehe dazu v. a. Hartshorne (2013), der überzeugend darlegt, dass es sich beim IC-Bias um eine grammatisch getriebene und nicht, wie in der Sozialpsychologie (zum Beispiel von Rudolph & Försterling 1997) angenommen, eine weltwissensbedingte Präferenz handelt.

Worin besteht nun diese Asymmetrie? Es ist prima facie nicht unbedingt einleuchtend, dass ein anaphorischer Prozess zwischen den thematischen Rollen Stimulus und Experiencer unterscheidet – üblicherweise wird angenommen, dass anaphorische Prozesse wie die Resolution eines (ambigen) Pronomens von Faktoren wie (linearer) Distanz, syntaktischer Bindung (im Falle von intrasententialen Vorkommen) und Diskursrelation (im Falle von intersententialen Anaphern) abhängen (siehe dazu z. B. Garnham & Cowles 2008; Kehler & Rohde 2013). Dies scheint zumindest für das Beispiel in (2) nicht der Fall zu sein: Lineare Distanz, syntaktische Gegebenheiten und Diskursrelation sind über die beiden Varianten hinweg konstant. Legt man die – in der Literatur mehr oder minder akzeptierte – Annahme zugrunde, dass die Anaphernresolution beim Sprachverstehen im Falle zweier möglicher Antezedenten denjenigen auswählt, dessen Repräsentation im Arbeitsgedächtnis der Hörerin/Leserin das höhere Aktivationsniveau hat, ergibt sich als Erklärung für den IC-Bias die folgende Salienzhypothese[3] (kurz: SH):

3 Wir sind uns im Klaren darüber, mit Salienz eine (kognitive) Größe in die Hypothese einzuführen, deren linguistische Reflexe und Rolle in der Verarbeitung notorisch

(SH) In einem Kontext der Form [NP1 Verb NP2, *weil* PersPron] wird das ambige Personalpronomen präferiert zugunsten der NP aufgelöst, die vom psychischen Verb die salientere Argumentrolle zugewiesen bekommt. Für psychische Verben ist dies unabhängig von der syntaktischen Funktion die Stimulus-Rolle.

Explizit vertreten wird die Salienzhypothese im Zusammenhang mit dem IC-Bias von Kasof & Lee (1993), wie das folgende Zitat illustriert:[4]

> Because people attribute greater causality to more salient stimuli than to less salient stimuli, people reading sentences implying different levels of salience for subjects and objects should attribute the interpersonal events unequally between subjects and objects. To the degree that a sentence evokes a mental representation in which the subject is more salient than the object, readers should attribute the interpersonal event more to the subject than to the object. To the degree that a sentence evokes a mental representation in which the object is more salient than the subject, readers should attribute the interpersonal event more to the object than to the subject. (ebd., p.878).

Diese Hypothese, die – mehr oder minder explizit – von den meisten Autoren, die zum IC-Bias arbeiten, geteilt wird, macht also eine Salienzasymmetrie auf der Ebene der Argumentrollen verantwortlich für die Abhängigkeit der Resolutionspräferenz für ambige Pronomina in *weil*-Fortsetzungen vom Typ des psychischen Verbs. Mit anderen Worten: Präsentiert man Probanden unvollständige Satzfolgen, bestehend aus einem Satz mit zwei genusidentischen NPen, die als Argumente eines psychischen Verbs fungieren, und einem mit *weil* und ambigem Personalpronomen eingeleiteten Satzfragment, so vervollständigen sie diese Satzfolgen präferent mit dem Stimulus-Argument des psychischen Verbs des ersten Satzes, weil das Stimulus-Argument salienter und damit das geeignetere Antezedens für das Personalpromonen ist.

unklar sind; dies geschieht zugegebenermaßen in Ermangelung eines weniger vagen Konzepts mit spezifischeren empirischen Korrelaten. Der Begriff der Prominenz scheint uns in diesem Zusammenhang nicht weniger problembehaftet. Siehe hierzu auch Vogel (2015).

4 *Stimulus* wird hier von den Autoren im Sinne von ‚in einem Experiment dargebotener Reiz', nicht im Sinne einer Theta-Rolle verwendet.

2.4 Zur Operationalisierung der Salienzhypothese im Korpus

Wie in Abschnitt 2.3 ausführlich dargestellt, wird in der Psycholinguistik als gängige Erklärung für den IC-Bias ein Salienzunterschied zwischen den Argumenten der in Rede stehenden Verben angesetzt. Experimentell wird das Salienzgefälle zwischen den Argumenten zumeist durch die Auswertung anaphorischer Bezüge ermittelt, wobei in der Regel angenommen wird, dass Probanden (beispielsweise in Satzvervollständigungsaufgaben) ein im nachfolgenden Satzfragment vorhandenes Pronomen auf das jeweils salientere Argument im Vorgängersatz beziehen. Insofern wäre es naheliegend, auch im Korpus anaphorische Bezüge auszuwerten und zu diesem Zwecke vorhandene Koreferenzannotationen zu nutzen. Abgesehen davon, dass Zweifel berechtigt sind, ob in den vorhandenen für Koreferenz annotierten Korpora überhaupt genügend Belege für die uns interessierenden psychischen Verben auffindbar wären, ist ein solches Vorgehen mit einem vergleichsweise hohen Aufwand verbunden. Die Korpora könnten nicht rein satzbezogen ausgewertet werden; vielmehr müsste für alle Fälle, in denen die untersuchten IC-Verben vorkommen, der jeweilige Kontext auf mögliche anaphorische Größen hin überprüft werden, die das Subjekt oder das Objekt des jeweiligen Verbs wiederaufnehmen. Dabei stellte sich nicht nur die Frage nach den berücksichtigten Formen (z. B. nur Pronomen oder auch komplexe NPen), sondern auch die Frage nach der Größe des Suchfensters: Wie viele Sätze können zwischen dem Zielsatz und dem Satz, der den anaphorischen Ausdruck enthält, liegen, damit der anaphorische Bezug für die Fragestellung noch relevant ist, und wie kann dieses über die Fälle hinweg normiert werden? Abgesehen von diesen Schwierigkeiten in der Umsetzung spricht gegen die Auswertung von anaphorischen Bezügen im Korpus aber vor allem, dass zur ökologischen Validierung der psycholinguistischen Salienzhypothese für den IC-Bias unabhängige Korpusevidenz nötig ist. Da das anaphorische Potenzial aber bereits in den psycholinguistischen Experimenten als abhängige Variable für die Ermittlung von Salienzunterschieden genutzt wurde, wäre dies mit der Auswertung von anaphorischen Bezügen über Koreferenzannotationen in Korpora nicht mehr ohne Weiteres gegeben. Vielmehr bedarf es eines anderen unabhängigen Maßes, um die experimentellen Ergebnisse korpusbasiert validieren zu können.

Welches Maß, das an der syntaktischen Oberfläche zugänglich ist, wäre nun geeignet, um Salienzunterschiede zwischen sprachlichen Ausdrücken im Korpus zu ermitteln? Salienz lässt sich im Korpus nur vermittelt und in Bezug auf grammatische Eigenschaften oder Relationen zwischen Ausdrücken feststellen, von denen unabhängig bekannt ist, dass sie die Salienz eines Ausdruckes bzw. genauer: seines Diskursreferenten erhöhen. Dies trifft beispielsweise auf Ersterwähnungen, Subjekte, Topiks usw. zu. Rein von der syntaktischen Oberfläche

aus gesehen sind diese funktionalen Ausdrücke aber unterschiedlich leicht ablesbar; am leichtesten ließen sich anhand der Kasusmarkierung wohl noch die Subjekte bestimmen. Um aber die Salienzhypothese zu überprüfen, bedarf es für die Korpusrecherche einer abhängigen Variable, die nicht nur einen nachvollziehbaren Schätzwert für die Salienz darstellt, sondern auch im Korpus beobachtbar, d. h. suchbar und zählbar ist. Passivkonstruktionen scheinen dieser Anforderung gerecht zu werden. Um die ökologische Validität der psycholinguistischen Salienzhypothese zu überprüfen, bedienen wir uns daher der passivierten Form von SE-und ES-Verben. Durch die Passivierung kann bei SE-Verben das Stimulus-Argument als optionales PP-Argument auftreten, während bei ES-Verben das Experiencer-Argument als optionales PP-Argument realisiert würde.[5] Nach Hypothese SH enthält diese optionale PP bei SE-Verben das salientere Argument, nicht jedoch bei ES-Verben. Diesen Zusammenhang nutzen wir für die Operationalisierung der Salienzhypothese im Rahmen unserer Korpusstudie. Wir nehmen an, dass die Realisierung des PP-Arguments im Passiv vom Salienzstatus des Referenten, der mit diesem Argument verbunden ist, abhängig ist, und erwarten, dass es für in diesem Sinne salientere Argumente wahrscheinlicher ist, als PP realisiert zu werden als für weniger saliente Argumente, weswegen Stimulus-Argumente häufiger als PP-Argument im Passiv vorkommen sollten als Experiencer-Argumente. Wird ein Satz mit einem SE-Verb passiviert (vgl. 3.a), sollten mehr overte PPen vorkommen als bei der Passivierung eines Satzes mit einem ES-Verb (vgl. 3.b).

(3) a. Erwin wird (von Siggi) geängstigt.
b. Siggi wird (von Erwin) gefürchtet.

Das Auftreten der PP im Passiv erscheint uns zum einen wegen seiner Optionalität und zum anderen wegen seiner problemlosen Überführbarkeit in Häufigkeiten als abhängige Variable besonders gut geeignet.

5 Die Frage nach der Motivation für die Passivierung blenden wir hier aus, auch weil es abgesehen von einem Verweis auf die mit der Subjektfunktion verbundene Prominenz und dem mit der Passivierung einhergehenden Perspektivenwechsel keine einfache Antwort darauf gibt. Was immer aber die Passivierung auslösen mag, diese Gründe bleiben konstant in Bezug auf die betrachteten Sätze mit oder ohne *von*-PP. Den Terminus *von*-PP verwenden wir hier als Oberbegriff für mit *von* und anderen, idiosynkratischen Präpositionen gebildeten PPen.

3 Korpusstudie

3.1 Design und Vorhersage

Überträgt man die aus der psycholinguistischen Literatur ableitbare Hypothese hinsichtlich des Salienzunterschiedes in Satzvervollständigungsexperimenten auf die hier interessierende Fragestellung hinsichtlich des Vorkommens der *von*-PP in Passivierungen, so lautet die empirische Vorhersage H_1 (komplementär zur Nullhypothese H_0):

(H_1): Stimulus-Argumente werden in Passivstrukturen häufiger overt in einer *von*-Phrase realisiert als Experiencer-Argumente, da erstere salienter sind als letztere.

Die zugehörige Nullhypothese lautet entsprechend, dass sich kein Unterschied in der Häufigkeit der overten Realisierung zwischen den beiden Argumenttypen findet. Die wichtigste unabhängige Variable war der zweistufige Faktor Verbtyp (SE- vs. ES-Verb); abhängige Variable war die relative Häufigkeit des Auftretens des im Passivsatz nicht als Subjekt realisierten Arguments in der *von*-Phrase. Dieses 1x2-Design wurde zusätzlich um den Zufallsfaktor Verbpaar erweitert, um eine Generalisierung über verschiedene SE/ES-Paare zu ermöglichen; Details hierzu werden im folgenden Abschnitt dargelegt.

3.2 Methode

Im Sinne der Zielvorgabe der Ermittlung der ökologischen Validität experimenteller Befunde zu IC-Verben wurden in einem ersten Schritt Verbpaare von Stimulus-Experiencer- und Experiencer-Stimulus-Verben identifiziert, für die experimentell erhobene Satzvervollständigungsdaten vorliegen und die über mehrere Experimente hinweg einen reliablen Unterschied gezeigt haben, der im Sinne der Salienzhypothese interpretierbar ist.[6] Das Kriterium der Paarbildung war dabei, dass eine Einsetzung konkreter Verben und ihrer Argumente in das folgende Schema eine plausible Beschreibung eines (kausalen) Zusammenhanges liefert:

6 Wir haben uns dabei an den Bias-Daten aus eigenen Satzvervollständigungsexperimenten sowie an den Bias-Daten von Ferstl, Garnham & Manoulidou (2011) zum Englischen orientiert. Dass der IC-Bias crosslinguistisch robust ist, zeigen die Arbeiten von Bott & Solstad (2014) und Hartshorne & Snedeker (2010).

(S): Wenn *x y* SE-verbt, dann ist es möglich, dass *y x* ES-verbt.

Das Paradebeispiel für eine Einsetzung in dieses Schema mit plausiblem Resultat ist das Paar *fürchten/ängstigen*: Wenn Peter Hans ängstigt, dann ist es möglich – und vielleicht sogar kausal erklärbar –, dass Hans Peter fürchtet. Dass sich dieses Schema nicht über alle Verbpaare gleich gut durchhalten ließ, zeigen Einsetzungsbeispiele wie *Wenn Peter Hans begeistert, dann ist es möglich, dass Hans Peter bewundert.* oder noch etwas weniger plausibel: *Wenn Peter Hans beunruhigt, dann ist es möglich, dass Hans Peter verdächtigt.* Entscheidend für diesen Test war letztlich, dass die Einsetzung des Verbpaares in das Schema nicht zu völliger Unplausibilität führt; das war für keines der ermittelten Verbpaare der Fall. Es ergaben sich die folgenden Verbpaare: *fürchten/ängstigen; bewundern/begeistern; bedauern/erschüttern; ertragen/stören; verachten/enttäuschen; hassen/beleidigen; bemerken/überraschen* und *verdächtigen/beunruhigen*. Die Aufgabe bei der Erstellung des Korpus war, für diese Verbpaare ein Korpus von je 200 Belegen von Passivsätzen (100 pro Verb) zu erstellen, was die Verwendung eines großen Korpus notwendig machte. Darüber hinaus sollten die Belege idealerweise innerhalb der Paare und über die Paare hinweg hinsichtlich ihrer Genrezugehörigkeit kontrolliert (d. h. *gematcht*) sein.

Extraktion der Stichprobe. Da DeReKo für die fraglichen Verbpaare im Passiv nicht hinreichend große Stichproben lieferte, haben wir das DWDS-Kernkorpus (vgl. Geyken 2007) genutzt, das gleichzeitig über ein Genrefeature verfügt. Wir verwendeten das über die URL <eins.dwds.de> zugängliche „alte" Graphic User Interface des DWDS, das die Belege nach Textsorten sortiert ausgibt, da dies eine kontrollierte Extraktion von Belegen in Abhängigkeit vom Genre ermöglichte. Die Suchanfrage lautete <"ge-V-t" @werden> – es wurden also alle Fälle des Partizips II mit adjazenter Form des Hilfsverbs extrahiert; darunter fanden sich einige Dubletten, Futur- und Konjunktivformen, die von Hand aussortiert werden mussten. Der Suche nach adjazenten Wörtern lag keine theoretische, sondern eine rein praktische Überlegung zugrunde, da die Suche nach nicht-adjazenten Wortformen (<"ge-V-t" && "werden">) zwar deutlich mehr Belege, aber eben auch mehr falschpositive Formen (z. B. Futur- und Konjunktivformen) zeitigte. Extrahiert wurden durch diese lemmabasierte Suchanfrage auch orthografisch abweichende Formen wie z. B. *geängstiget wardt*. Tabelle 1 gibt einen Überblick über die Anzahl von Belegen.

Unser Bestreben, innerhalb der Verbpaare eine möglichst balancierte Verteilung hinsichtlich des Faktors Genre zu haben, war leider nicht bei allen Verbpaaren erfolgreich. Für Verbpaarlinge mit wenigen Belegen (d. h. etwa 120 Fälle, davon im Mittel ca. 20 Falschpositive) waren wir auf die Genrezusammensetzung dieses Verbs festgelegt; der andere Paarling musste dann, soweit möglich, der Genrezusammensetzung des ersten Paarlings angepasst werden. Abbildung 1

Tabelle 1: Anzahl extrahierter Fälle von Passivformen nach Verbtyp.

SE-Verben	N		ES-Verben	N
ängstigen	109		fürchten	307
begeistern	145		bewundern	872
erschüttern	1.758		bedauern	175
stören	3.056		ertragen	386
enttäuschen	1.592		verachten	403
beleidigen	626		hassen	111
überraschen	2.192		bemerken	973
beunruhigen	118		verdächtigen	411
TOTAL	9.596			3.638

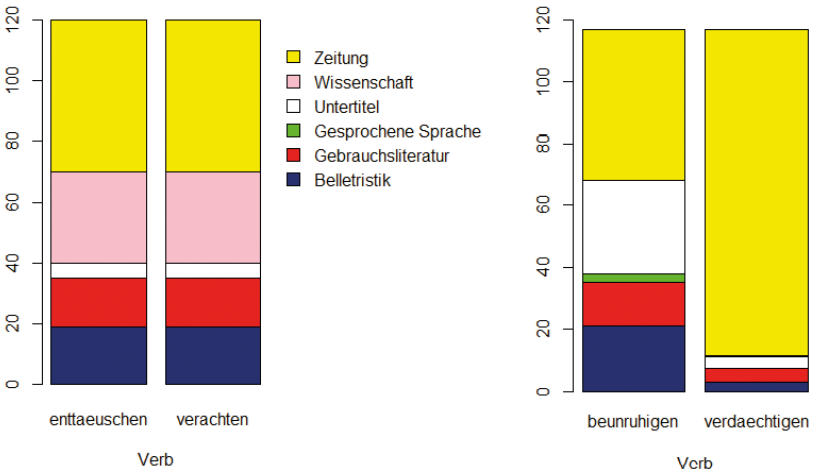


Abbildung 1: Genrezusammensetzung für die Verbpaare *enttäuschen/verachten* und *beunruhigen/verdächtigen*.

zeigt zwei extreme Fälle, die durch dieses Vorgehen entstanden sind: Das Verbpaar *enttäuschen/verachten* ließ sich hinsichtlich Genre vollständig balancieren; die maximale Abweichung zwischen zwei Paarlingen mussten wir beim Verbpaar *beunruhigen/verdächtigen* in Kauf nehmen.

In einigen wenigen Fällen mussten nach Bereinigung der extrahierten 120 Belege noch Belege „nachgezogen" werden, da zu viele Dubletten und Falschpositive enthalten waren. Nach diesem Schritt lagen für jeden der 16 Verbpaarlinge 100 Passivformen vor.

Datenaufbereitung. Die extrahierten und bereinigten 1.600 Fälle wurden mithilfe des Statistik-Software-Pakets R (R Core Team 2014, Version 3.1.1) so aufbereitet, dass jeder der Belege, die in ihrer Länge und Komplexität massiv variierten, in einer Tabelle einheitlich wie folgt repräsentiert war: Die Partizipialform bildete eine eigene Spalte, der diese Form umgebende Text (soweit vorhanden) die linke und rechte Spalte daneben. Vor diese Spalten wurde eine Spalte mit einer durchlaufenden Zählvariable, eine Spalte mit dem Verb und eine Spalte mit der Ausprägung des Faktors VERBTYP gesetzt. Hinter der Spalte mit dem Text, der der Partizipialform folgte, wurden die zu annotierenden Merkmale jeweils spaltenweise angeordnet.

Annotation. Die Annotation erfolgte in MS Excel anhand folgender Merkmale:[7]

- VERBTYPE: SE- vs. ES-Verb (automatisch erzeugt)
- PASSIV: Passivform vs. keine Passivform ([0,1]-kodiert)
- VON.PP: Realisierung eines der beiden Argumente in einer *von*-Phrase (oder einer äquivalenten PP; [0,1]-kodiert, wie auch alle folgenden Merkmale)
- IDIO.PP: idiosynkratische, vom Verb selegierte Präpositionen (z.B. *geängstigt werden durch*; [0,1]-kodiert)
- IDIO.PP.FORM: Wortform der idiosynkratischen PP
- S.ANIM: Belebtheitsstatus des Subjekts
- PP.OBJ.ANIM: Belebtheitsstatus des PP-Objekts
- S.DEF: Definitheitsstatus des Subjekts
- PP.OBJ.DEF: Definitheitsstatus des PP-Objekts
- POSITION: relative Position von Subjekt und PP-Objekt; die unmarkierte Abfolge wurde als „0“ kodiert.

Die Annotation des Merkmals, das für unsere Fragestellung zentral ist, VON.PP, erwies sich als vollständig unproblematisch. Hier zwei Beispiele für Kodierung mit „1“ und „0“; im Folgenden wird die Verbform unterstrichen und die *von*-PP **in Fettdruck** wiedergegeben:

(4) Auch in diesem Jahr dürfte der Kreisverkehr, der selbst **von versierten Automobilisten** gefürchtet wird, seine Spitzenstellung halten – wenn nicht bald etwas geschieht.[8]

7 Hier möchten wir für die Unterstützung bei der Annotation der Daten Miriam Feix, Cheryl Hodgkinson, Heinke Jank und Markus Paluch danken.

8 DWDS, Subkorpus *Berliner Zeitung*; Textklasse: Zeitung::Lokales; Berliner Zeitung vom 01.07.1997.

(5) Zum 13. Mal erzielte Polster in einer Bundesligapartie zwei Treffer – gefürchtet wird er unter dem Künstlernamen Toni Doppelpack.[9]

Auch die Annotation des Merkmals IDIO.PP war, von einigen wenigen Zweifelsfällen abgesehen, unproblematisch; diese Zweifelsfälle – handelt es sich bei einer idiosynkratischen PP tatsächlich um das Argument, das dem Subjekt im Aktivsatz entspricht? – wurden im Kreis der Annotatoren diskutiert und geklärt. Hier ein Beispiel für IDIO.PP==1 und IDIO.PP.FORM==*mit*:

(6) Ist der Darwinismus des Marktes, **mit dem** wir täglich geängstigt werden, nur das Spiel einer Elite?[10]

Größere Schwierigkeiten bereitete – wenig überraschend – die Annotation der Belebtheitsmerkmale S.ANIM und PP.ANIM. In Zweifelsfällen wie NPen, die Gremien, Institutionen und sonstige Kollektive von Individuen bezeichnen *(der Aufsichtsrat, das Parlament, die Studierendenschaft*) optierten wir im Zweifel für den Merkmalswert „belebt", da es die individuellen psychischen Zustände und Absichten der in die Summenindividuen eingehenden Elemente sind (die Aufsichtsräte, die Parlamentarier, die Studierenden), die Träger der EXPERIENCER- bzw. STIMULUS-Eigenschaften sind. Ein Beispiel für einen unzweifelhaften Fall von S.ANIM==0 gibt (7):

(7) In der Regel bezieht sich die fixe Idee auf unerreichte Zwecke, auf Güter, die gehofft, auf Uebel, die gefürchtet werden.[11]

Der Definitheitsstatus der beiden Argumente bereitete keine größeren Annotationsprobleme. Allerdings traten Fälle von koordinierten NPen auf, die gemischten Belebtheits- oder Definitheitsstatus aufwiesen; hier ein Beispiel für letzteren:

(8) So haben wir gesehen, daß der große Komet des Jahres 1456
Entsetzen über ganz Europa verbreitete, das ohnehin schon **durch**

9 DWDS, Subkorpus *Berliner Zeitung*; Textklasse: Zeitung::Sport; Berliner Zeitung vom 26.08.1996.

10 DWDS, Subkorpus *Berliner Zeitung*; Textklasse: Zeitung::Magazin; Berliner Zeitung vom 06.12.1997.

11 DWDS, Subkorpus *Deutsches Textarchiv*; Textklasse: Wissenschaft::Medizin; Reil, Johann Christian (1803), Rhapsodieen über die Anwendung der psychischen Curmethode auf Geisteszerrüttungen. Halle.

eine verheerende Pest und **durch die Verwüstungen, welche die Türken um sich verbreiteten**, geängstiget wurde.[12]

In Fällen wie (8) wurde das Merkmal PP.DEF mit „9“ kodiert, was in der statistischen Analyse als fehlender Wert interpretiert wurde.

Linguistisch interessanter waren Problemfälle, die sich aus der Split-Stimulus-Konstruktion (siehe Engelberg 2015) ergaben; (9) gibt ein Beispiel:

(9) Die Ausschreitungen von damals seien „bis heute für Rostock ein Brandmal“, sagte Gauck, dessen Rede kurzzeitig **von Zwischenrufen wie „Heuchler“ durch Linksautonome** gestört wurde.

Anhand von (9) lassen sich einige der Annotationsprobleme nochmals illustrieren: Das Subjekt des passivierten Relativsatzes ist das EXPERIENCER-Argument *dessen* [also Gaucks] *Rede*; dies wurde folglich als S.ANIM==0 und S.DEF==1 annotiert. Das STIMULUS-Argument allerdings ist zweigeteilt. Gaucks Rede wird *von* Zwischenrufen *durch* Linksautonome gestört, d. h. die VON-PP ist unbelebt, die IDIO.PP aber belebt.
Da die Wortstellung von Subjekt und *von*-PP im Satz höchst selten markiert war, verursachte auch diese Annotation keine Probleme.

Zur Erinnerung: Wir sagen, der Salienzhypothese folgend, vorher, dass sich die abhängige Variable – relative Auftretenshäufigkeit der VON-PP bzw. IDIO-PP – in Abhängigkeit vom Faktor VERBTYP unterschiedlich verhalten sollte. Die Fälle ohne overte PP sollten im Falle passivierter ES-Verben häufiger sein als im Falle von SE-Verben, weil bei letzteren das (per Hypothese salientere) STIMULUS-Argument der Kandidat für die Realisierung der PP ist. Und umgekehrt, und aus demselben Grund, sollten die Fälle von overt realisierter PP häufiger bei SE- als bei ES-Verben auftreten. Statistisch sagen wir also eine (disordinale) Interaktion der Faktoren VERBTYP und REALISIERUNG DER PP vorher.

12 DWDS, Subkorpus *Deutsches Textarchiv*; Textklasse: Gebrauchsliteratur::Populärwissenschaft::Wissenschaft::Physik; Littrow, Joseph Johann von (1836): Die Wunder des Himmels, oder gemeinfaßliche Darstellung des Weltsystems. Bd. 3. Stuttgart.

3.3 Ergebnisse

Die annotierten 1.600 Fälle verteilten sich wie folgt auf die sich aus der Kreuzung der beiden Faktoren ergebenden vier Zellen:

Tabelle 2: Absolute Häufigkeiten der Realisierung der PP in Abhängigkeit vom Faktor VERBTYP.

	Verbtyp	
	ES	SE
ohne overte PP	583	408
mit overter PP	217	392

Abbildung 2 illustriert dieses Befundmuster grafisch.

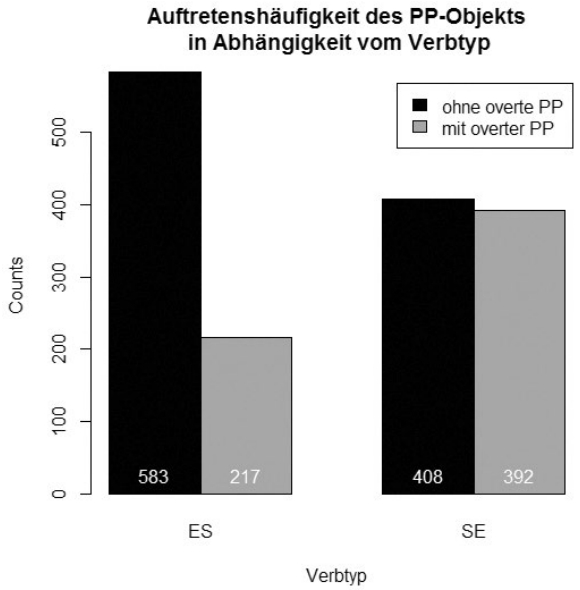


Abbildung 2: Absolute Häufigkeiten von Passivsätzen mit overter und ohne overte PP in Abhängigkeit vom Verbtyp.

Um das sich numerisch andeutende Interaktionsmuster – 466 Fälle Unterschied zwischen overter vs. non-overter PP für ES-Verben und nur 16 Fälle Unterschied für SE-Verben – inferenzstatistisch zu erhärten, wurde eine *mixed-model* logistische Regression auf der [0,1]-kodierten abhängigen Variablen *Realisierung der PP* gerechnet; diese ergab sich aus dem Aufaddieren der mit „1“ kodierten Fälle von

realisierten *von*-PPen und idiosynkratischen PPen.[13] Die Berechnung der logistischen Regression wurde in R mit dem Paket *lme4* (Version 3.1, Bates et al. 2015) und dem Befehl für generalisierte lineare Modelle, *glmer*, durchgeführt. Die acht Verbpaare behandelten wir dabei als Zufallsfaktor *(random factor)* und VERBTYP als festen (*fixed factor)*.

Dieses Verfahren schätzt den Einfluss des Prädiktors VERBTYP auf die logit-transformierte abhängige Variable *Realisierung der PP*. Das heißt, es gibt uns eine Antwort auf die Frage, wie viel der im Datensatz vorhandenen Varianz dem manipulierten Faktor zuzuschreiben ist und wie viel davon reine Fehlervarianz („Rauschen") ist. Die Fehlervarianz kann bei dem von uns gewählten Verfahren nochmals durch die Berücksichtigung des Messwiederholungsfaktors VERBPAAR unterteilt werden: Ist der Einfluss des Faktors VERBTYP signifikant (d. h. lässt er eine Generalisierung auf die „Population" von Sätzen zu, aus der wir unsere Stichprobe gezogen haben), so sollte sich dieser Einfluss innerhalb der Verbpaare und über diese hinweg zeigen. Die logit-Transformation der abhängigen Variablen wird dabei durchgeführt, um eine α-Fehler-Inflation zu vermeiden, die bei der Berechnung des Modells auf untransformierten absoluten Häufigkeiten droht (siehe Agresti 2002). Ein signifikantes Ergebnis für eine logistische Regression besagt in unserem Fall also, kurz gesagt, dass wir die Nullhypothese, dass der untersuchte Faktor VERBTYP keinen Einfluss auf die Auftretenswahrscheinlichkeit der PP hat, verwerfen und die Alternativhypothese (VERBTYP hat einen Einfluss) annehmen dürfen. Der *p*-Wert gibt uns dabei die Wahrscheinlichkeit an, mit der wir ein der Alternativhypothese entsprechendes Datenmuster (oder ein extremeres) finden können, obwohl in der Population das der Nullhypothese entsprechende Datenmuster gilt. Wir folgen den geltenden Standards und nehmen ein α-Fehler-Niveau von .05 an; der ermittelte *p*-Wert sollte also unter diesem kriterialen Wert liegen.[14]

Die logistische Regression ergab einen signifikanten Einfluss des Faktors VERBTYP auf die Häufigkeit der Realisierung der PP (β= 1.01, $|z|$=2.76, p = .04) – der Einfluss des Faktors VERBTYP auf die Gesamtvarianz im Datensatz ist also reliabel nachweisbar: Das Nicht-Subjekt-Argument eines passivierten psychischen Verbs tritt reliabel häufiger auf, wenn es sich um ein STIMULUS-Argument als wenn es sich um ein EXPERIENCER-Argument handelt.

13 Ein Pearson-*chi*^{2}-Test ist bei diesem Design wegen der Abhängigkeit der Faktorstufen nicht möglich.

14 Die Modellgleichung der *mixed model* logistischen Regression lautete wie folgt: glmer(PPA~verbtype+(1+verbtype|item), data=d.clean, family=binomial, REML=-FALSE). Dieses Modell wurde mit dem Nullmodel (d.h. PPA~1+(1+verbtype|item) …) verglichen; vgl. zu diesem Vorgehen Barr et al. (2013). Weitere Details der statistischen Analyse werden hier aus Platzgründen ausgespart; Rohdaten und R-Skripte stellen wir auf Anfrage gern zur Verfügung.

Post-hoc-Analyse. Nimmt man allerdings weitere Prädiktoren wie beispielsweise die Definitheit des Subjekts und dessen Belebtheit ins Modell auf, über deren Einfluss wir keine Hypothese formuliert haben und die damit als *post-hoc*-Faktoren anzusehen sind, so sinkt der Beitrag von Verbtyp zur Varianzaufklärung in den marginal signifikanten Bereich (p = .08 bei Hinzunahme von S.Anim als Faktor und p = .11 bei Hinzunahme von S.Anim und S.Def.)[15]

Das bedeutet, dass man zur Vorhersage der Häufigkeit der Realisierung einer *von*-PP in einem Satz mit passiviertem psychischen Verb den Faktor Verbtyp zwar heranziehen kann, und eine relativ gute Vorhersage des Auftretens der PP innerhalb der Verbpaare und über die Verbpaare hinweg erhält. Ein guter Teil dieser Vorhersagekraft dieses Faktors speist sich aber offenbar aus anderen Faktoren, deren wirkmächtigste der Belebtheitsstatus und die Definitheit des Subjekts sind.

Dies wirft die Frage auf, wie sich die abhängige Variable aus unserer Korpusstudie zu den aus psycholinguistischen Satzvervollständigungsexperimenten stammenden IC-Bias-Werten verhält. Da beide Variable [0,1]-kodiert sind, lassen sie sich ohne Transformation auf einer Skala abtragen; Abbildung 3 stellt den direkten Vergleich der Daten aus den beiden Evidenzquellen grafisch dar. Wie man anhand der unterschiedlichen Steigungen der Geraden in Abbildung 3 sieht, ist der experimentell erhobene Bias stärker als seine Entsprechung in der Korpusstudie.

In einem weiteren Schritt kann man den Zusammenhang zwischen den Korpus- und den experimentellen Daten auf Ebene der einzelnen Verbpaare betrachten; dies soll Abbildung 4 leisten:

Wie Abbildung 4 zeigt, ist der experimentell erhobene Bias über die Verbpaare hinweg relativ stabil. Der Effekt des Faktors Verbtyp variiert auf dieser Variablen zwar in seiner Stärke (d. h. der Steigung der Geraden, die die beiden Mittelwerte verbindet), aber das Vorzeichen der Steigung ist stets positiv. Das gilt für die Korpusdaten nicht. Hier gibt es zwei Verbpaare mit negativer Steigung auf der abhängigen Variablen (*hassen/beleidigen* und *verachten/enttäuschen*) sowie ein Verbpaar nahezu ohne Effekt, d. h. einer Geraden ohne Steigung *(bewundern/begeistern)*. Durch unsere *post-hoc*-Tests haben wir mindestens zwei der Faktoren (Definitheits- und Belebtheitsstatus des Subjekts) identifiziert, denen wir diese Abweichung zwischen den beiden Evidenzquellen zuschreiben können.

15 Die Hinzunahme des Belebtheits- und Definitheitsstatus des PP-Objekts führte nicht zu einer signifikanten Veränderung der Varianzaufklärung. Im Modellvergleich (Likelihood Ratio chi^2-Test, siehe Barr et al. 2013), dem derzeit wahrscheinlich konservativsten Test für gemischte Modelle, liegt der p-Wert für den Vergleich des komplexesten Modells – mit Verbtyp, S.Anim und S.Def – mit dem Vergleichsmodell – mit S.Anim und S.Def – bei p =.12.

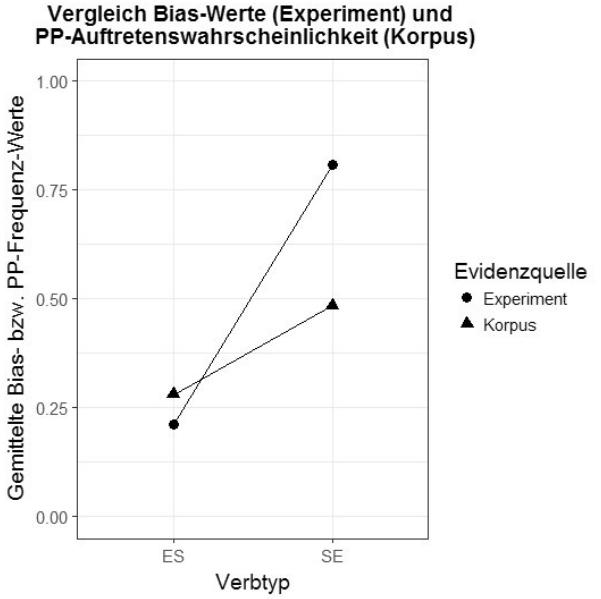


Abbildung 3: Vergleich der Mittelwerte von experimentell erhobenem Anaphernresolutions-Bias (●) und der Auftretenswahrscheinlichkeit der PP in unserem Passivkorpus (▲).

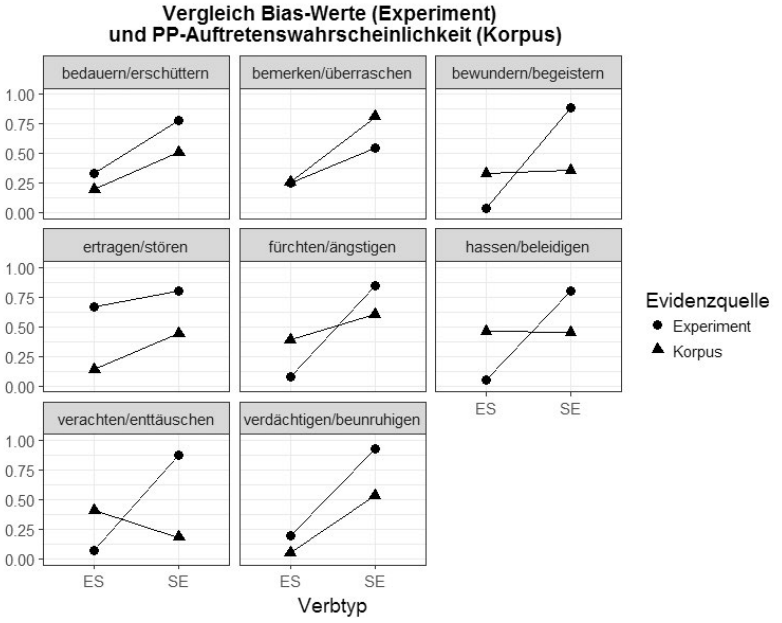


Abbildung 4: Vergleich der Mittelwerte von experimentell erhobenem Anaphernresolutions-Bias (●) und der Auftretenswahrscheinlichkeit der PP in unserem Passivkorpus (▲).

Selbstverständlich wären weitere Einflussgrößen denkbar, die die Abweichung zwischen Experiment und Korpus erklärbar machen, insbesondere informationsstrukturelle und diskursbezogene Faktoren. Diese zu erfassen hätte allerdings den ohnehin beträchtlichen Annotationsaufwand noch vergrößert, und ihre Einbeziehung in eine statististische Analyse des Auftretens der *von*-PP in Passivsätzen mit Psychverben bleibt zukünftigen Untersuchungen vorbehalten.

4 Diskussion

Die Befunde der logistischen Regression für den Faktor haben gezeigt, dass der üblicherweise in psycholinguistischen Experimenten manipulierte Faktor, die Zugehörigkeit des Verbs zur SE- vs. zur ES-Klasse, die Auftretenswahrscheinlichkeit von PPen in Sätzen mit passivierten psychischen Verben verlässlich vorhersagen kann. Allerdings zeigten die *post-hoc*-Analysen, dass der Einfluss dieses Faktors nicht unabhängig zu sehen ist von Eigenschaften der Argumente, die in Satzvervollständigungsexperimenten üblicherweise *nicht* manipuliert werden: Der Definitheits- und der Belebtheitsstatus des Subjekts sind starke Prädiktoren für das Auftreten der *von*-PP – und sie sind in ihrem Potenzial zur Aufklärung von Varianz in den Korpusdaten deutlich stärker als der Faktor Verbtyp. In psycholinguistischen (Satzvervollständigungs-)Experimenten werden zumeist Eigennamen als Argumente der Verben dargeboten, d. h. belebte definite Ausdrücke. Eine Ausnahme ist in dieser Hinsicht die Studie von Corrigan (1988), die den Belebtheitsstatus der Argumente in drei Experimenten systematisch mit dem Faktor Verbtyp gekreuzt hat und signifikante Interaktionen dieser Faktoren sowohl für Kausalitätsratings als auch für Satzvervollständigungen gefunden hat. Während bei SE-Verben auch inanimaten Referenten von Stimulus-Argumenten eine kausale Rolle im vom Verb denotierten Ereignis attribuiert wird, ist dies bei ES-Verben in weitaus geringerem Ausmaß der Fall. Und auf unbelebte Stimulus-Referenten von SE-Verben wird in Satzvervollständigungen häufiger Bezug genommen als auf die Referenten unbelebter Stimuli von ES-Verben. Worauf Corrigans Daten hindeuten, ist eine Korrelation zwischen verbspezifischen Belebtheitsrestriktionen und den abhängigen Variablen typischer psycholinguistischer Experimente zu IC-Verben – eine Korrelation, die von den meisten nachfolgenden Studien ignoriert wurde.[16] Unsere Korpusdaten deuten in eine ähnliche Richtung. Während SE-Verben offenbar weniger anfällig für den Effekt der Belebtheit des Subjekts (d. h., im Passiv, des Stimulus-Argumentes) sind, hat

16 So stellen Kasof & Lee schon 1993 fest: „Prior research on implicit causality has centered on a rather narrow range of sentence forms." (ebd., S. 878). Leider hat sich daran auch seitdem nicht viel geändert.

die Animatheit des Subjekts auf die ES-Verben einen deutlich stärkeren Einfluss. Die stärkere Streuung, die die verschiedenen SE-Verben in Abbildung 4 hinsichtlich der Korpusdaten aufweisen, lässt sich – tentativ und *post hoc* – über den Einfluss der verbspezifischen Animatheitsanforderung bei diesen Verben erklären: je weniger stark die Belebtheitsanforderung an das Stimulus-Argument, desto niedriger der Score der Korpusdaten. Dies sei an einem Beispiel illustriert. Das Stimulus-Argument eines STÖREN- oder ÜBERRASCHEN-Ereignisses ist hinsichtlich seines Belebtheitsstatus möglicherweise weniger festgelegt als ein ENTTÄUSCHEN- oder BELEIDIGEN-Ereignis. Während der Stimulus des Gestörtwerdens ebenso gut ein Geräusch sein kann wie eine Person, die das Geräusch produziert, ist es schwerer vorstellbar, dass einem Beleidigtwerden ein unbelebter Stimulus zugrunde liegt. Eine systematische quantitative Analyse des Einflusses dieser Animatheitsprofile (also: animat-animat, inanimat-inanimat, animat-inanimat, inanimat-animat) der Verben ist aufgrund der Ungleichverteilung der Profile in unserem Korpus nicht möglich. Es scheint aber lohnend, solche Analysen anhand größerer Korpora zu erstellen, um diese Faktoren und ihr Zusammenspiel in balancierten Subkorpora zu untersuchen. Wir können somit festhalten, dass der Belebtheitsstatus eine geringere Rolle spielt, wenn das Stimulus-Argument im Aktiv als Subjekt des Satzes realisiert wird.

Damit schließt die Diskussion unmittelbar an die Frage nach der ökologischen Validität der experimentell erhobenen Daten zum IC-Bias an. Diese ist offenbar eher niedrig, weil die Faktoren, die in den psycholinguistischen Experimenten im Regelfall konstant gehalten werden, de facto einen großen Einfluss auf die beobachtete Argumentasymmetrie der hier untersuchten psychischen Verben haben können. Der Preis des kontrollierten Vorgehens beim Experiment, welches notwendig zum Ausschluss möglichst vieler denkbarer Störvariablen führt, besteht also im Informationsverlust über ebenfalls vorhandene grammatische Einflussgrößen, die die Ausprägung der abhängigen Variablen potenziell modulieren. Insofern lässt sich Schütze (2006) nur beipflichten, wenn er schreibt: „Because no single kind of data is perfect, an efficacious approach to linguistic investigation is to seek converging evidence from a wide array of types of data whenever possible." Die hier vorgestellte Fallstudie sehen wir als weiteren Beleg für unsere Überzeugung, dass empirische Studien in der Linguistik an Aussagekraft gewinnen, wenn sie auf konvergenten Ergebnissen beruhen, die durch verschiedene Methoden erzielt wurden. Dies eröffnet zugleich die Möglichkeit, Korpora zur Validierung experimenteller Daten zu nutzen. Während experimentelle Studien kontrolliert sind und daher immer nur Ausschnitte des jeweils untersuchten Phänomens beleuchten können, repräsentieren Korpora das gesamte Spektrum der möglichen Einflussgrößen, allerdings konfundiert, d. h. in Korpusanalysen können diese Faktoren nicht kontrolliert werden. Das ermöglicht es aber, Korpora zur ökologischen Validierung und damit ggf. als Korrektiv

für experimentell erhobene Ergebnisse heranzuziehen. Darüber hinaus eröffnet diese Strategie die Möglichkeit, im Korpus identifizierte Einflussgrößen – wie in unserem Fall Animatheit und Definitheit – als Faktoren in kontrollierte Experimente aufzunehmen und dort einer systematischen, d. h. nicht-konfundierten Untersuchung ihres Effekts zuzuführen.

Zusammenfassend lässt sich konstatieren, dass die beiden betrachteten Methoden – Korpusanalyse und Experiment – gerade wegen der Komplementarität ihrer jeweiligen Stärken und Schwächen letztlich immer aufeinander bezogen und idealerweise systematisch parallelisiert durchgeführt werden sollten.

Literaturverzeichnis

Agresti, Alan (2002): Categorical Data Analysis. Hoboken, NJ: Wiley.

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

Bates, Douglas/Mächler, Martin/Bolker, Benjamin M./Walker, Steven C. (2015): Fitting Linear Mixed Effects Models Using lme4. In: Journal of Statistical Software 67(1), doi: 10.18637/jss.v067.i01.

Belletti, Adriana/Rizzi, Luigi (1988): Psych Verbs and Theta Theory. Natural Language and Linguistic Theory 6, S. 291–352.

Bott, Oliver/Solstad, Torgim (2014): From Verbs to Discourse: A novel account of implicit causality. In: Hemforth, Barbara/Mertins, Barbara/Fabricius-Hansen, Cathrine (Hg.): Meaning and Understanding across Languages. Studies in theoretical psycholinguistics. Chicago: Springer International, S. 213–251.

Brown, Roger/Fish, Deborah (1983): The psychological causality implicit in language. In: Cognition 14, S. 237–273.

Corrigan, Roberta (1988): Who dun it? The influence of actor-patient animacy and type of verb in the making of causal attribution. In: Journal of Memory and Language 27, S. 447–465.

Engelberg, Stefan (2015): Gespaltene Stimulus-Argumente bei Psych-Verben. Quantitative Verteilungsdaten als Indikator für die Dynamik sprachlichen Wissens über Argumentstrukturen. In: Engelberg, Stefan/Meliss, Meike/Proost, Kristel/Winkler, Edeltraud (Hg.): Argumentstruktur – Valenz – Konstruktionen. Tübingen: Narr, S. 469–491.

Ferstl, Evelyn/Garnham, Alan/Manouilidou, Christina (2011): Implicit causality bias in English: a corpus of 300 verbs. In: Behavior Research Methods 43(1), S. 124–135.

Garvey, Cathrine/Caramazza, Alfonso (1974): Implicit causality in verbs. In: Linguistic Inquiry 5, S. 459–464.

Garnham, Alan/Wind Cowles, Heidi (2008): Looking both ways: The JANUS model of noun phrase anaphor processing. In: Gundel, Jeanette K./Hedberg, Nancy (Hg.): Reference: Interdisciplinary perspectives. Oxford: Oxford University Press, S. 246–272.

Hartshorne, Joshua K. (2013): What is implicit causality? In: Language, Cognition and Neuroscience 29(7), S. 804–824.

Kasof, Joseph/Lee, Ju Young (1993): Implicit Causality as Implicit Salience. In: Journal of Personality and Social Psychology 65(5), S. 877–891.

Kehler, Andrew/Rohde, Hannah (2013): A Probabilistic Reconciliation of Coherence-Driven and Centering-Driven Theories of Pronoun Interpretation. In: Theoretical Linguistics 39, S. 1–37.

Pawlik, Kurt (1976): Ökologische Validität: Ein Beispiel aus der Kulturvergleichsforschung. In: Kaminski, G. (Hg.): Umweltpsychologie. Perspektiven – Probleme – Praxis. Stuttgart: Klett, S. 59–72.

Pickering, Martin J./Majid, Asifa (2007): What are implicit causality and implicit consequentiality? In: Language and Cognitive Processes 22, S. 780–788.

Postal, Paul M. (1971): Cross-Over Phenomena. New York: Holt, Rinehart and Winston, S. 39ff.

R Core Team (2014): R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. URL http://www.R-project.org/.

Schütze, Carston T. (2006): Data and Evidence. In: Brown, Keith (Hg.): Encyclopedia of Language and Linguistics. 2. Aufl., Bd. 3, Oxford: Elsevier, S. 356–363.

Vogel, Ralf (2015): Is prominence a useful concept for the theory of syntax? Lingue e Linguaggio, 14(1), 97-112. doi:10.1418/80756.

Wirtz, Markus Antonius (Hg.) (2013): Dorsch – Lexikon der Psychologie. 16. Aufl., Verlag Hans Huber.

Markus Bader, Vasiliki Koukoulioti

When Object-Subject Order is Preferred to Subject-Object Order: The Case of German Main and Relative Clauses[1]

Abstract Overall, subject-before-object (SO) order is preferred in German to object-before-subject order (OS), as reflected in higher acceptability and higher frequency of the former in comparison to the latter. Certain conditions have been identified, however, where OS order is preferred to SO order. First, main clauses in which the object is related to the prior discourse by a partially-ordered set relation, and second, relative clauses with a personal pronoun as subject. In order to explore the circumstances under which OS is preferred to SO order, we present preliminary data from ongoing corpus studies investigating relative clauses and main clauses in which either the subject or the object occupies the prefield. The corpus data confirm prior findings from experimental studies and extend them in several ways. In particular, the corpus data reveal a close connection between referential form and word order, with demonstrative pronouns strongly favoring the use of OS order.

Keywords German syntax, word order, prefield, relative clauses, referential form, language production, topic

1 Introduction

Although German is considered as a language with relatively free word-order, sentences in which the subject precedes the object(s) occur with a much higher frequency than sentences in which the subject follows one or more objects (Hoberg 1981; Kempen & Harbusch 2005; Bader & Häussler 2010). Not at least for this reason subject-before-object order (SO) is generally considered to be the

1 This work was supported by the Deutsche Forschungsgemeinschaft (Project VER within the Research Unit 1783 "Relative Clauses").We would like to thank Yvonne Portele, Alice Schäfer and an anonymous reviewer for their helpful comments.

canonical order of subject and object in German. Although clauses deviating from the canonical order – that is, clauses with object-before-subject order (OS) – are much less frequent when considered across the board, under certain circumstances, OS order seems to be preferred to SO order.

Many studies on language processing across a variety of languages have found that sentences with non-canonical argument order are more difficult in comparison to sentences with canonical argument order. Sentences with OS order and passive sentences are acquired later than sentences with SO order (e.g., Friedmann et al. 2009), they are less often produced than sentences with SO order (e.g., Bader & Häussler 2010), they pose severe problems for people with aphasia (e.g., Burchert et al. 2008), and even for adult speakers without any language disturbance, they are often more difficult to comprehend than corresponding SO sentences (e.g., Kaiser & Trueswell 2004).

In some cases, however, the disadvantage for non-canonical sentences vanishes or is even reversed to an advantage. Weskott, Hörnig, Fanselow & Kliegl (2011) coined the terms *weak* and *strong licensing* of the OS order for such cases. Weak licensing refers to the situation where the SO and the OS variant of a sentence are equivalent with regard to measures like acceptability and processing complexity. Strong OS licensing, on the other hand, obtains when the OS order is at an advantage in comparison to the SO variant. In the following, we will focus on corpus frequencies as an indicator of strong or weak OS licensing, but other measures will also be taken into account if available.

A prominent case of strong OS licensing identified by corpus linguistic studies of German word order is illustrated by the two sentences in (1), which differ only with regard to the order of subject and object.

(1) a. Wahrscheinlich wird der Fehler dem Lehrer entgehen.
likely will the teacher the error miss
'The teacher will probably miss the error.'

b. Wahrscheinlich wird dem Lehrer der Fehler entgehen.
likely will the teacher the error miss
'The teacher will probably miss the error.'

The sentences in (1) contain a non-agentive verb, an inanimate subject and an animate object. Given this particular configuration of verb semantics and animacy of the arguments, sentences with OS order occur with higher frequency than sentences with SO order (Hoberg 1981; Bader & Häussler 2010; Verhoeven 2015), and they receive higher ratings in acceptability experiments (Ellsiepen & Bader 2018).

In addition to verb-semantics and animacy, which together comprise the class of lexical-conceptual factors, factors concerning the discourse status of

the individual NPs are known to affect the choice between SO and OS order. In comparison to the lexical-conceptual factors discussed above, discourse-related factors have received less attention in corpus studies on German word order, in particular with regard to the issue of strong and weak OS licensing. However, as will be discussed below, at least two cases have been identified in the experimental literature, one concerning relative clauses and one concerning main clauses with either the subject or the object occupying the prefield.

Because main clauses and relative clauses differ in many ways, they may seem like an odd pair as far as the choice between SO and OS order is concerned. The most important difference in the current context concerns the degree of optionality with regard to the order of subject and object. Since it is obligatory to front relative pronouns in German, OS order can be obligatory for relative clauses in some cases, as in the example in (2).

(2) Das ist der Lehrer, dem das Buch gefallen hat.
this is the teacher who the book pleased has
'This is the teacher who the book pleased.'

Given the meaning expressed by sentence (2), the relative clause must occur with OS order. The object relative pronoun must occur clause-initially and a verb like *gefallen* ('to please') cannot be passivized. It is therefore not possible to turn the dative object into a subject, thereby producing a subject-initial relative clause instead of an object-initial one. In this respect, sentences as in (2) contrast with sentences containing a relative clause that allows passivization, as shown in (3). Here, instead of producing an object-initial relative clause with the verb in the active voice, a subject-initial relative clause with a verb in the passive voice can be produced as an alternative.

(3) a. Das ist der Lehrer, den der Schüler gegrüßt hat.
this is the teacher who the student greeted has
'This is the teacher who the student greeted.'
b. Das ist der Lehrer, der von dem Schüler gegrüßt wurde.
this is the teacher who by the student greeted was
'This is the teacher who was greeted by the student.'

In main clauses, in contrast, there is almost always a choice between putting the subject or the object into the prefield. In (4), for example, the same truth-conditional meaning can be expressed either with SO order (4a) or with OS order (4b).

(4) a. Peter hat diesen Film schon zweimal gesehen.
P. has this movie already twice seen
'Peter has already seen this movie twice.'
b. Diesen Film hat Peter schon zweimal gesehen.
this movie has P. already twice seen
'Peter has already seen this movie twice.'

To say that the order of subject and object is optional in (4) is not to say that the two orders can be freely exchanged in all contexts. Quite to the contrary, it is a truism that in most cases of word order optionality, the alternative orders are associated with different usage conditions. Starting with the seminal work of Lenerz (1977) and Höhle (1982), the pragmatic conditions that license the use of SO or OS order have been the topic of extensive research. The major insight emanating from this research is that SO sentences are typically (relatively) unrestricted with regard to discourse conditions, allowing uses with both wide and narrow focus, whereas OS sentences typically require narrow focus on one of their constituents. The exact conditions vary depending on whether both subject and object are contained within the middle field or whether one has been moved to the prefield, as in the examples in (4) (see Frey 2004b for the relationship between these two cases). In the following, we will consider only main clauses with either subject or object in the prefield. The question of whether to choose SO or OS order therefore boils down to the question of whether to move the subject or the object to the prefield.

This paper presents data from three ongoing corpus studies, one investigating relative clauses and two investigating main clauses. The reason for investigating these two clause types relates to the issue of strong versus weak OS licensing. Two questions will be pursued in this regard: first, can the experimental findings concerning the conditions that weakly or strongly license the use of OS order be replicated when looking at written language production, and second, can the set of conditions leading to weak or strong OS licensing be extended? Relative clauses are discussed in more detail in the next section. Afterwards, we turn to main clauses. In the final section, we discuss the implications of our findings for future research.

2 Non-canonical order in relative clauses

Relative clauses have played a major role in research on language acquisition and language processing, both disturbed and undisturbed. A common finding of this research is that subject-initial relative clauses as in (5a) are easier than object-initial relative clauses as in (5b), for both children and adults.

(5) a. The gardener who$_i$ t$_i$ contacted the reporter left early.
b. The gardener who$_i$ the reporter t$_i$ saw left early.

Starting with Fox and Thompson (1990), it has become clear that object relatives are not in general more difficult than subject relatives. Based on an investigation of naturally occurring relative clauses, Fox and Thompson (1990) showed that object relative clauses occur particularly often with a personal pronoun as subject. Thus, in contrast to relative clauses in which the second NP is a lexical NP, as in (5), object relatives prevail when the second NP is a pronoun, as in the following example.

(6) a. The gardener who$_i$ t$_i$ contacted me left early.
b. The gardener who$_i$ I contacted t$_i$ left early.

Later research has extended this finding to language comprehension (Mak et al. 2008) and to language acquisition (Kidd et al. 2007). For both English and German child language, Kidd et al. (2007) present corpus counts as well as experimental evidence showing that the large majority of object relative clauses produced by children has a pronoun as subject.

What has not been shown so far is whether the same also holds for adult German language production. In an ongoing corpus study of written German relative clauses, we are currently analyzing a set of about 1700 relative clauses randomly drawn from the deWac corpus (Baroni et al. 2009). In this paper, we present selected preliminary results concerning the distribution of the referential form of the second NP in subject and object relative clauses.

644 relative clauses contained both a subject and a direct object with either one being realized as relative pronoun. Of these, 547 relative clauses (85 %) were subject-initial and 97 relative clauses (5 %) were object-initial. Overall, subject relatives clearly outweigh object relatives. We classified the second argument of each relative clause – that is, the object in subject relatives and the subject in object relatives – with regard to the type of NP, using the same categories as Kidd et al. (2007): first-person pronoun, second-person pronoun, third-person pronoun, proper name, lexical NP, and others. We used one additional category not used by Kidd et al., namely reflexive pronouns. Table 1 shows the distribution of the relative clauses according to the NP type of the second NP, depending on whether the relative clause occurred with SO or OS order.

In subject relative clauses, the second NP is a lexical NP most of the time. Reflexives also occur with some regularity, whereas all other categories are quite rare. For object relatives, in contrast, the second NP is a personal pronoun in the majority of cases, with third- and first- person pronouns occurring most often and with about equal frequency. Lexical NPs also appear as second NP

Table 1: Percentages of relative clauses with different types of the second NP, depending on the syntactic function of the relative pronoun. Raw numbers are given in parentheses.

	Lexical NP	Reflexive	3ps pro	Proper Name	1ps pro	2ps pro	Other
Subj. relative	71.8 (392)	19.8 (107)	3.5 (19)	0.7 (4)	0.4 (2)	0.0 (0)	3.9 (23)
Obj. relative	17.5 (17)	0.0 (0)	28.9 (28)	1.0 (1)	32.0 (31)	5.2 (5)	15.5 (15)

in object relative clauses, but with a strikingly lower frequency than in subject relatives.

For object relatives, our results are similar to those of Kidd et al. (2007) (results for subject relatives are not reported by them). The major difference is that in our study, third- and first-person pronouns occur almost equally often whereas the majority of pronouns in Kidd et al.'s study were first-person pronouns. This difference can be attributed to the fact that we analyzed a corpus sample of written adult language whereas Kidd et al. analyzed a corpus of spoken child language.

Mak et al. (2008) have proposed that the subject in an object relative clause is typically a topic, and that this explains the high proportion of subject pronouns in object relatives. If the subject is a topic even when it is not a pronoun, this makes the prediction that the subject should immediately follow the relative pronoun in most cases because the leftmost position within the middlefield is the default position for topics (Frey 2004a). For subject relatives, it is assumed that the object is typically not a topic. Objects in subject relatives are therefore expected to occur anywhere within the relative clause. To test this hypothesis, we determined the clausal position in which the second NP appears for each relative clause. This is the subject when the relative pronoun is the object and it is the object when the relative pronoun is the subject. The relative pronoun always occurs in position 1. If the second NP occurs directly after the relative pronoun, it appears in position 2. If exactly one phrase intervenes between relative pronoun and second NP, the second NP occurs in position 3, and so on. In our corpus sample, the second NP occurred in one of positions 2–6.

Table 2 shows how often the second NP occurs in each position for both subject and object relatives. Relative clauses in which the second NP was either a personal or a reflexive pronoun were excluded from this analysis because such pronouns obligatorily occur in an early position within the clause. In object relatives, the second NP appears in position 2 in nearly all cases, that is, directly after the relative pronoun. Because the second NP in an object relative clause is the subject, it typically occupies a high position in the syntactic structure, therefore occurring rather early in the clause. However, only specific, topical subjects appear in the highest position below the prefield (Diesing 1992; Frey

Table 2: Percentages of relative clauses in which the second NP (the object in subject relative clauses, the subject in object relative clauses) occurs in clausal positions 2-6, where position 1 is the relative pronoun. Relative clauses where NP2 was either a personal pronoun or a reflexive were excluded from the analysis. Raw numbers are given in parentheses.

	2	3	4	5	6
Subject relative	63.6 (335)	25.3 (134)	8.2 (45)	1.9 (10)	0.9 (1)
Object relative	93.3 (28)	3.3 (1)	3.3 (1)	0 (0)	0 (0)

2004a), and the finding that the subject in object relative clauses occurs in the second position in over 90 % of all cases is therefore suggestive for these subjects being topics. Because the exact position of the subject often remains ambiguous unless there is an adverbial marking the left VP boundary, the evidence is only suggestive and in need of further confirmation. In subject relatives, position 2 is also the most frequent position for the second NP, but here later positions are also observed with some regularity.

In sum, we take the data shown in Table 2 as tentative support of Mak et al.'s claim that the second NP is a topic in an object relative clause but not in a subject relative clause. Note that the frequency counts shown in Table 2 cannot be reduced to the definiteness of the second NP. Definite NPs are known to occur earlier in the clause than indefinite NPs (Lenerz 1977; for corpus evidence, see Bader & Häussler 2010). However, in both subject and object relative clauses, the second NP was a definite NP in the majority of cases, with no significant difference between the two clause types.

A further finding concerning the use of object relatives has been reported by Contemori and Belletti (2014). In an experiment investigating the spoken production of Italian relative clauses, Contemori and Belletti found that adults have a strong preference for producing subject relatives with the verb in the passive voice instead of corresponding object relatives. For children, such a preference became visible only in the oldest age group investigated (between 8 and 9 years).

In order to test whether a similar preference holds in the corpus sample under consideration, we determined the number of relative clauses with a verb in the passive voice and with or without a *von* ('by') PP. Since we are only considering verbs with an accusative object in this paper, these are all subject relative clauses. Table 3 shows the resulting numbers as well as the number of relative clauses with a verb in the active voice, a subject and an accusative object (these are the same subject and object relatives already discussed above).

Table 3 shows that the overall frequency of passive subject relative clauses is higher than the frequency of object relative clauses. However, a clear majority

Table 3: Number of subject and object relative clauses with a verb in the active voice, a subject and an accusative object, and number of subject relative clauses with a verb in the passive voice and with or without a *von* ('by') PP.

	Active	Passive with *von* PP	Passive without *von* PP
Subject relative clause	547	18	136
Object relative clause	97	–	–

of passive relative clauses does not contain a *von*–PP. Thus, passive voice seems to be used in relative clauses mainly for the purpose of omitting the underlying subject, and not for the purpose of avoiding OS order. The numbers in Table 3 contrast with the experimental results of Contemori and Belletti (2014). It is an open question as to whether this is related to grammatical differences between Italian and German or to other differences (e.g., spoken versus written language, experimental data versus corpus data).

In sum, the corpus data presented in this section reveal the same pattern as has been found for English child and adult language and for German child language. Whereas subject relatives outnumber object relatives when considering all relative clauses with a subject and a direct object, the reverse relationship is found when we look only at relative clauses in which the second NP is a personal pronoun. Here, object relatives occur with greater frequency than subject relatives, especially for first and second person pronouns. The additional data discussed in this section support Mak et al.'s (2008) hypothesis that the high percentage of subject pronouns in object relatives comes about because the subject in an object relative clause is a topic. On the other hand, we found no evidence that writers revert to the passive voice in order to avoid object relative clauses.

3 Non-canonical order in main clauses

We now turn to main clauses in which either the subject or the object occupies the prefield. As before, we restrict our discussion to sentences with an accusative object. For main clauses, the situation is more complex than for relative clauses because there is no constraint restricting the clause-initial phrase in a way similar to the case of relative clauses. With regard to factors favoring OS order, discourse properties of the object can therefore be as relevant as discourse properties of the subject. Before we consider subject and object in turn, the next subsection reviews recent theories of how speakers or writers decide which phrase to put into the prefield.

The preferred filler of the prefield

The question of whether to put the subject or the object into the prefield is closely related to the question of what the preferred position of the sentence topic is in German main clauses. While older work saw the prefield as the default position for the sentence topic (Gundel 1988), more recent research (Frey 2004a; Rambow 1993; Speyer 2007) suggests that the default position is at the left edge of the middlefield (sometimes called the Wackernagel position). For reasons of space, we consider only the proposal of Speyer (2007, 2009, 2010), who claims that the topic is put into the prefield only if a clause contains no element higher on the prefield hierarchy given in (7).

(7) scene-setting >> poset >> topic

The hierarchy in (7) contains two kinds of elements that have precedence when it comes to filling the prefield. Scene-setting elements are typically adverbials that locate an event in time and space. A poset element is linked to the prior discourse by a poset (partially ordered set) relation in the sense of Ward and Prince (Ward & Prince 1991). Examples for poset relations are the set-membership relation and the part-of relation.

For purposes of illustration, we consider the experiments of Weskott et al. (2011), which confirm the importance of the poset relation for the purposes of filling the prefield. Weskott et al. (2011) obtained acceptability ratings and reading times for short texts consisting of two sentences, as illustrated in (8).

(8) Peter hat den Wagen gewaschen.
Peter has the.ACC car washed.
'Peter has washed the car'

a. Er hat den Außenspiegel ausgelassen.
He.NOM has the.ACC side mirror left-out
b. Den Außenspiegel hat er ausgelassen.
The.ACC side mirror has he.NOM left-out.
'The side mirror, he left out.'

The first sentence of each text introduces two referents. The first of them is taken up again in the second sentence by a subject pronoun. The second referent of the initial sentence is not taken up again in toto, but a part of it is referred to in the second sentence by means of a definite NP serving as the object. Thus, the NP *den Außenspiegel* ('the side mirror') in (8) stands in a poset relation to the NP *den Wagen* ('the car') of the first sentence. As shown in (8), the second sentence

appears with either SO or OS order. Both acceptability ratings and reading times revealed an advantage for OS sentences in comparison to SO sentences. This is therefore a case of strong OS licensing. Following Speyer (2007), Weskott et al. (2011) attribute the strong OS licensing for sentences as in (8) to the poset relation between the object of the second and the object of the first sentence, but they note that taking up the first NP by means of a subject pronoun may also have contributed to the advantage observed for OS order.

In the next two subsections, we present corpus data addressing two questions. The first is whether the referential form of the subject affects the probability of using a sentence with OS order. In particular, does a pronominal subject have a similar effect as seen in relative clauses? The second question concerns the object itself. Given that Weskott et al. (2011) found strong licensing when the object referent stood in a poset relation to a prior referent, the question is whether other relationships between the object referent and the prior discourse strongly or weakly license OS order as well. Here, we will consider the simplest relation, namely the identity relation which holds when the object referent is simply given in the prior discourse.

Subject properties favoring OS order

In accordance with prior findings on child language and language comprehension, our corpus study of relative clauses found that in the majority of object relative clauses the second NP is a personal pronoun. Whether we should expect a similar finding for main clauses is not straightforward because word order is less optional in relative clauses than in main clauses. That is, whereas declarative main clauses leave a choice as to which element to put into the prefield, there is no choice of word order in relative clauses as far as the initial element is concerned – the relative pronoun always has to come first.

Preliminary evidence on this issue comes from an ongoing corpus study that investigates the conditions governing the choice between personal pronoun and d-pronoun. This is a follow-up study to Portele and Bader's (2016) study, which investigated the choice between personal pronoun and d-pronoun for the case of subject pronouns. Based on a search of about 20 % of the deWac Corpus (Baroni et al. 2009), Table 4 shows how the form of the NP immediately following the finite verb in a verb-second clause depends on properties of the phrase that fills the prefield.

When the prefield is filled by a subject pronoun, the percentage of personal pronouns is quite low. When the prefield is filled by an object pronoun, in contrast, personal pronouns are found much more often directly after the finite verb. This is so when the personal pronoun *ihn* fills the prefield, and to an even greater

Table 4: Percentages of personal pronouns and other elements directly following the finite verb in a verb-second clause depending on syntactic function and the pronoun type of the pronoun in the prefield.

Syntactic function	Pronoun type	Word form	Element after C°	
			Personal pronoun	Other
Subject	Personal pronoun	Er	5.7	94.2
	D-pronoun	Der	8.1	91.9
Direct object	Personal pronoun	Ihn	22.3	77.8
	D-pronoun	Den	59.9	40.1

extent when the prefield hosts the d-pronoun *den*. In fact, almost 60% of all sentences starting with the d-pronoun *den* had a personal pronoun as the subject. Sentences with a d-pronoun in the middlefield are contained within the category of other elements after C° in Table 4. A preliminary analysis of this category revealed that d-pronouns occur quite infrequently as subjects within the middlefield, a finding which has also been obtained by Bosch, Katz and Umbach (2007). Thus, sentences in which the object is a d-pronoun and the subject a personal pronoun seem to constitute a further case of strong OS licensing.

In sum, in both relative and main clauses the probability of OS order increases when the subject is a pronoun. In contrast to relative clauses, the evidence for main clauses is only suggestive. Further research is necessary – including research on language acquisition and language comprehension – in order to determine how far the parallels go.

Object properties favoring OS order

The corpus data presented in this section are from an ongoing corpus study testing the prefield hierarchy given in (7) for the case of sentences with a subject and an accusative object. This study analyzes sentences from a random selection of Wikipedia texts, including 10,000 Wikipedia articles for each letter of the alphabet unless fewer were available. Because the focus is on determining what properties of the object increase or decrease its probability of occurring before or after the topic of the sentence, all sentences analyzed in this study contain the subject pronoun *er* ('he') with the discourse function of topic. Five types of object NPs are analyzed:[2]

2 The corpus search also included the d-pronoun *den* ('the.ACC') but there were too few corpus hits to warrant further analysis.

— definite NPs starting with the definite article *den* ('the.ACC')
— indefinite NPs starting with the indefinite article *einen* ('a.ACC')
— demonstrative NPs starting with the demonstrative determiner *diesen* ('this.ACC')
— the personal pronoun *ihn* ('him.ACC')
— the demonstrative pronoun *diesen* ('this.ACC')

Table 5 shows the number of corpus hits for each of the five types of object NPs listed above. With regard to the proportion of OS sentences, Table 5 shows a clear distinction depending on the type of the object NP. When the object is either a definite NP, an indefinite NP or a personal pronoun, SO order is preferred. This preference is strongest in the case of personal pronouns, which is in accordance with linguistic descriptions according to which object pronouns cannot occupy the prefield except for highly specific discourse conditions, including contrastive stress (see Lenerz 1992). For demonstrative objects, however, a preference for OS order is observed, for both full and pronominal NPs.

Table 5: Number of corpus hits broken down by order and type of the object NP.

	Non-pronominal object			Pronominal object	
	Demonstrative NP	Definite NP	Indefinite NP	Demonstrative pronoun	Personal pronoun
SO	121	3860	1907	35	183
OS	388	841	305	110	4
% OS	76.2	17.9	13.9	75.8	2.1
Ratio	1 : 3.2	4.6 : 1	6.3 : 1	1 : 3.1	45.8 : 1

Definite NPs in particular, but indefinite NPs to some degree too, are known to show a great variety of relationships to the prior discourse. In order to investigate how the relation of non-pronominal objects to the prior discourse affects word order, 100 corpus examples with a non-pronominal object for each of the six combinations of order and object type were randomly selected for a detailed analysis. With regard to the discourse status of the object NP, the examples were annotated using the classification proposed in Birner and Ward (2009), which extends the influential proposal of Prince (1981). According to Birner and Ward (2009), each referent can be classified as given or new in two dimensions. The first dimension concerns the prior discourse: a referent can have been mentioned in the prior discourse or it can have been newly introduced. The second dimension concerns the hearer (or reader): a referent can be old or new relative to the hearer's prior knowledge. Birner and Ward make the further assumption that the two dimensions can be freely combined, giving rise to four categories as shown

Table 6: Discourse status as proposed by Birner & Ward (2009) following Prince (1981).

	Hearer-old	Hearer-new
Discourse-old	Evoked *he*	Inferable
Discourse-new	Unused *Gov. Rod Blagojevich*	Brandnew *a $3.6 billion state construction budget*

in Table 6. The names for the four categories are from Prince (1981). An example is given in (9). The bold-printed phrases in (9) illustrate three of the four categories shown in Table 6.

(9) **Gov. Rod Blagojevich**, while scaling back a massive capital program, said Friday **he** would endorse **a $3.6 billion state construction budget** that includes new money to build schools and millions of dollars for legislative pork-barrel projects.
(Chicago Tribune, 8/23/03) [from (Birner 2003)]

The referent of the proper name ***Gov. Rod Blagojevich*** is mentioned for the first time at this point of the discourse and is therefore discourse-new. Since the typical reader of the newspaper where this text is from can be assumed to be familiar with this referent, it is hearer-old. This referent is taken up again by the following personal pronoun *he*, which thus refers to a referent evoked in the preceding clause and is therefore both discourse- and hearer-old. The indefinite NP *a $3.6 billion state construction budget* introduces a new referent that cannot be assumed to be known by a typical reader. This referent is thus brand-new, that is, discourse- and hearer-new. Example (9) does not contain an instance of the fourth category, the inferables. However, we already saw an example of an inferable referent when discussing the experiments of Weskott et al. (2011). In example (8), the referent of the NP *den Außenspiegel* ('the side mirror') has not been mentioned before, but it can be inferred because it stands in a poset relation (more precisely a part-of relation) to the car mentioned in the preceding sentence.

We next discuss the main findings for each of the three types of non-pronominal NPs included in the present corpus study. For each NP type, a representative example containing a main clause with OS order is provided in Table 7.

(i) *Demonstrative objects.* Demonstrative objects show a rather uniform relationship to the preceding context. In almost all cases, they refer to a referent evoked in the immediately preceding clause. Most of the time, this is the referent of an NP, as in the example in Table 7, but references to the event

Table 7: Representative examples of OS sentences in which the fronted object was either a demonstrative NP, an indefinite NP, or a definite NP.

Demonstrative NP	Alfred Alexander Taylor (. . .) war ein US-amerikanischer Politiker und der 38. Gouverneur von Tennessee. *Diesen Bundesstaat* vertrat er außerdem im US-Repräsentantenhaus.
	'Alfred Alexander Taylor (. . .) was an US-American politician and the 38th Governor of Tennessee. He also represented *this state* in the House of Representatives.'
Indefinite NP	Anschließend promovierte Monar im Jahre 1989 an der Ludwig-Maximilians-Universität München in moderner Geschichte. *Einen zweiten Doktortitel* erlangte er im Jahre 1991 auf dem Gebiet der Politik- und Sozialwissenschaften am Europäischen Hochschulinstitut in Florenz.
	'Subsequently, Monar graduated in 1989 in modern history from Ludwig-Maximilians-University of Munich. A second doctoral degree he achieved in 1991 in the field of political and social sciences from the University of Florence.'
Definite NP	Loos starb im Sanatorium Kalksburg bei Wien, wo er mit einer Krankenschwester befreundet war, die er dem Vernehmen nach heiraten wollte. Er ruht in einem Grab auf dem Wiener Zentralfriedhof (Gruppe 0, Reihe 1, Nummer 105). *Den Grabstein* hatte er selbst entworfen.
	'Loos died in the Kalksburg sanatory near Vienna, where he was friends with a nurse, who he wanted to marry, it is said. He rests in a grave at the Vienna Central Cemetery (Group 0, Row 1, Number 105). *The gravestone* he had designed himself.'

introduced by the VP are also not uncommon (e.g., *Last year, Peter won the German championship. This victory …*). As shown in Table 5, demonstrative objects occur more often in the prefield than in the middlefield and thus constitute an instance of strong OS licensing. This is not predicted by the prefield hierarchy of Speyer because the relevant discourse relation – identity with a referent evoked in the prior discourse without being a topic – does not appear in the prefield hierarchy.

(ii) *Indefinite objects.* Indefinite NPs introducing a brand-new referent occur in the middlefield most of the time. When the referent of indefinite NPs stands in a poset relation to a referent in the prior discourse, as in the example in Table 7, the indefinite NP preferentially appears in the prefield. A similar observation has been made for English by Ward and Prince (1991).

(iii) *Definite objects.* As expected given the linguistic literature, definite NPs showed the most varied behavior in terms of discourse status. A preference for the prefield and thus OS order was only found for definite NPs in a poset relation to the prior discourse. For NPs which were inferable from the situation as a whole, but not from a specific referent in the prior discourse, in contrast, SO order prevailed (see Ward & Prince, 1991, for the

difference between NPs given by a poset relation and NPs that are situationally given). Anaphoric definite NPs, that is, NPs referring to referents that are discourse- and hearer-old, are not uncommon in the prefield, but they are even more common in the middlefield. Definite NPs referring to an unused referent, that is, a referent that is discourse-new but hearer-old, appear most of the time in the middlefield.

Toward a prefield hierarchy for SO/OS order

The findings reviewed in this section are mostly compatible with Speyer's prefield hierarchy in (7). Only one discrepancy was found: NPs referring to given referents show a preference for the prefield if the NP is a demonstrative – either an NP or pronoun – or a d-pronoun. For the task of choosing between subject and object as the filler of the prefield, we therefore propose the following prefield hierarchy.

(10) SO/OS prefield hierarchy
given(demonstrative, d-pronoun), poset > topic, given(definite) > brand-new

Like the more general prefield hierarchy of Speyer, the SO/OS prefield hierarchy in (10) is not meant as a categorical hierarchy but a preference hierarchy which captures preferences in the case of competing orders. Note that the SO/OS prefield hierarchy differs from Speyer's prefield hierarchy not only with regard to the number of elements, but also with regard to the type of information that is referred to. In contrast to Speyer's hierarchy, the hierarchy in (10) refers not only to the discourse status of the various referents but also to the referential form used for making reference. The finding that given referents which are referred to by a demonstrative expression are especially prone to fill the prefield and thus to occur in sentence initial position may possibly be related to the very nature of demonstratives, that is, pointing to an element in the nearby context (see Consten and Averintseva-Klisch 2010).

4 General discussion

As noted in the introduction, sentences with non-canonical word order are typically acquired later and are more difficult to process than sentences with canonical word order. However, in some cases, sentences with non-canonical word order are in fact advantageous in comparison to sentences with canonical word

order, as captured in Weskott et al.'s (2011) notion of strong licensing of OS order. The first question asked in this paper was whether reported instances of strong OS licensing can be replicated when looking at written language production in German. The second question was whether additional instances of strong (or weak) OS licensing can be found.

With respect to relative clauses, we found that object relative clauses are produced more frequently in written language when the second NP is a topic, an entity already introduced in the discourse. This claim is based on two findings. First, we found that object relative clauses are more frequent when the second NP is a pronoun, which typically refers to topics. A second finding was that in object relative clauses the subject almost always occurs directly after the relative pronoun, which is the canonical topic position.

These corpus findings are in accordance with previous experimental findings. For Dutch, Mak et al. (2008) report that object relative clauses are easier to process than subject relative clauses when the second NP is a case-ambiguous pronoun. This means that when a relative clause is processed, a pronoun that does not commit the reader to a specific reading is preferably interpreted as the subject of the relative clause and consequently the relative clause is interpreted as an object relative clause. Our findings replicate this comprehension pattern in written language production. Mak et al. (2008) also manipulated the context of subject and object relative clauses. They presented subject and object relative clauses in neutral and topic contexts (which introduced the second NP of the relative clause). They found that when the second NP was introduced and thus the topic, object relative clauses were equally easy (but not easier) to process as subject relative clauses.

The corpus analysis of relative clauses showed that passive voice in subject relative clauses occurs most of the time without a by-phrase. This is in contrast to the findings of Contemori and Belletti (2014). However, this pattern resembles the findings of Friedmann et al. (2009), who report that Hebrew-speaking children in some cases produced subject relative clauses with a reflexive verb instead of an object relative clause. In any case, our corpus findings suggest that passive subject relative clauses are not used as an alternative to object active clauses.

All in all, our findings confirm earlier findings that object relative clauses are not less frequent than subject relative clauses across the board, but are in fact preferred under specific conditions related to discourse factors. In particular, the present findings provide further evidence for strong OS licensing when the subject of a relative clause is a topic, and especially so when it is a pronominal topic. This may also explain why in the study of Hirschberg et al. (2014) object relatives occur with a rather high percentage of about 25%. Although subject relatives (which are not differentiated with regard to whether they also contain an object) are still the most frequent type in this study, the percentage of object

relatives is much higher than in our study or in Mak et al. (2002). Hirschberg et al. (2014) investigate a corpus of spoken language and almost all examples of object relatives contain a first-person pronoun as subject. As shown above, this is exactly the condition that strongly favors the production of object relatives.

With respect to main clauses, the situation is more complicated because the order of subject and object is affected by properties of both constituents. First, we found that – similarly to relative clauses – the proportion of OS sentences increases when the subject is a personal pronoun. Since this finding was restricted to sentences in which the object is a pronoun, further corpus research is necessary to determine whether this finding generalizes to other types of object NPs. With respect to the object, we found that word order is affected both by the relation of the object to the prior discourse and by the particular referential expression of the object NP (demonstrative vs. definite vs. indefinite NP vs. personal pronoun).

The corpus data discussed in this paper raise a range of questions in need of further research. First, in contrast to experimental research on relative clauses, few experimental studies exist on strong OS licensing for main clauses. For acceptability ratings and reading times obtained for adult participants, Weskott et al. (2011) have shown strong OS licensing when the object is related by a poset relation to the prior discourse, but for the other cases experimental evidence is lacking (for related work on language acquisition, see Sauermann 2016).

If the cases of strong OS licensing in main clauses can be corroborated, a further question is whether the findings can be accommodated within an overarching account. In particular, main clauses and relative clauses were similar insofar as the object precedes the subject more frequently when the subject is the topic than when the subject is not a topic. Can this similarity be rooted in the discourse function associated with topics, or is this just a superficial similarity between relative clauses and main clauses that has no common source?

A final set of questions concerns the SO/OS prefield hierarchy proposed in (10). One task for future research is to integrate the SO/OS prefield hierarchy, which only applies to the order of subject and object, with Speyer's hierarchy, which applies to all potential fillers of the prefield. To do so, it has to be determined, among others, how scene-setting phrases and demonstrative NPs are ranked relative to each other. In addition, it remains to be seen how lexical-conceptual information (e.g., animacy and thematic roles) interacts with the discourse-based information encoded in the SO/OS prefield hierarchy. As shown by several corpus studies (e.g., Bader & Häussler 2010; Verhoeven 2015), lexical-conceptual information does not only affect the order of arguments within the middlefield but also when one argument occupies the prefield. Addressing this issue will require taking prefield and middlefield into account simultaneously (see Frey 2004b for a theoretical-linguistic proposal.)

References

Bader, Markus and Jana Häussler. 2010. Word order in German: A corpus study. *Lingua* 120(3): 717–762.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi and Eros Zanchetta. 2009. The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation Journal* 23(3): 209–226.

Birner, Betty. 2003. Discourse functions at the periphery: Noncanonical word order in English. In B. Shaer, Werner Frey and Claudia Maienborn (eds.), *Proceedings of Dislocated Elements Workshop, ZAS Berlin*, 41–62.

Birner, Betty J. and Gregory Ward. 2009. Information structure and syntactic structure. *Language and Linguistics Compass* 3(4): 1167–1187.

Bosch, Peter, Graham Katz and Carla Umbach. 2007. The non-subject bias of German. In Monika Schwarz-Friesel, Manfred Consten and Mareile Knees (eds.), *Anaphors in Text: Cognitive, formal and applied approaches to anaphoric reference*, 145–164. Amsterdam: John Benjamins Publishing.

Burchert, Frank, Nadine Meißner and Ria d. Bleser. 2008. Production of non-canonical sentences in agrammatic aphasia: Limits in representation or rule application? *Brain and language* 104(2): 170–179.

Consten, Manfred and Maria Averintseva-Klisch. 2010. 'Nahe Referenten'-ein integrativer Ansatz zur Funktion demonstrativer Referenz. *Sprachtheorie und germanistische Linguistik* 20(1): 1–34.

Contemori, Carla and Adriana Belletti. 2014. Relatives and passive object relatives in Italian-speaking children and adults: Intervention in production and comprehension. *Applied Psycholinguistics* 35(06): 1021–1053.

Diesing, Molly. 1992. *Indefinites*. Cambridge, MA: MIT Press.

Ellsiepen, Emilia and Markus Bader. 2018. Constraints on argument linearization in German. *Glossa.*: *Glossa: a journal of general linguistics*, *3*(1), 6. DOI: http://doi.org/10.5334/gjgl.258.

Fox, Barbara A. and Sandra A. Thompson. 1990. A discourse explanation of the grammar of relative clauses in English discourses. *Language* 66: 297–316.

Frey, Werner. 2004a. A medial topic position for German. *Linguistische Berichte* 198: 153–190.

Frey, Werner. 2004b. The grammar-pragmatics interface and the German prefield. *Sprache & Pragmatik* 52: 1–39.

Friedmann, Naama, Adriana Belletti and Luigi Rizzi. 2009. Relativized relatives: Types of intervention in the acquisition of A-bar dependencies. *Lingua* 119(1): 67–88.

Gundel, Jeanette K. 1988. Universals of topic-comment structure. In Michael Hammond, Edith A. Moravcsik and Jessica R. Wirth (eds.), *Studies in syntactic typology*, 209–239. Amsterdam: John Benjamins Publishing.

Hirschberg, Tim, Carolin Reinert, Anna Roth and Caroline Féry. 2014. Relative Clauses in Colloquial and Literary German: A Contrastive Corpus-Based Study. *Linguistische Berichte* 240: 405–445.

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

Höhle, Tilman N. 1982. Explikation für „normale Betonung" und „normale Wortstellung". In Werner Abraham (ed.), *Satzglieder im Deutschen. Vorschläge zur syntaktischen, semantischen und pragmatischen Fundierung*, 75–153. Tübingen: Narr.

Kaiser, Elsi and John C. Trueswell. 2004. The role of discourse context in the processing of a flexible word-order language. *Cognition* 94: 113–147.

Kempen, Gerard and Karin Harbusch. 2005. The relationship between grammaticality ratings and corpus frequencies: A case study into word-order variability in the midfield of German clauses. In Marga Reis and Stephan Kepser (eds.), *Linguistic evidence. Empirical, theoretical and computational perspectives*, 329–349. Berlin: de Gruyter.

Kidd, Evan, Silke Brandt, Elena Lieven and Michael Tomasello. 2007. Object relatives made easy: A cross-linguistic comparison of the constraints influencing young children's processing of relative clauses. *Language and cognitive processes* 22(6): 860–897.

Lenerz, Jürgen. 1977. *Zur Abfolge nominaler Satzglieder im Deutschen*. Tübingen: Narr.

Lenerz, Jürgen. 1992. Zur Syntax der Pronomina im Deutschen. *Sprache und Pragmatik* 29.

Mak, Willem, Wietske Vonk and Herbert Schriefers. 2002. The influence of animacy on relative clause processing. *Journal of Memory and Language* 47: 50–68.

Mak, Willem M., Wietske Vonk and Herbert Schriefers. 2008. Discourse structure and relative clause processing. *Memory & Cognition* 36(1): 170–181.

Portele, Yvonne and Markus Bader. 2016. Accessibility and referential choice: Personal pronouns and d-pronouns in written German. *Discours. Revue de linguistique, psycholinguistique et informatique* 18: 1–41.

Prince, Ellen F. 1981. Toward a taxonomy of given-new information. In Peter Cole (ed.), *Radical pragmatics*, 223–255. New York etc.: Academic Press.

Rambow, Owen. 1993. Pragmatic aspects of scrambling and topicalization in German: A Centering Approach. In *IRCS Workshop on Centering in Discourse*.

Sauermann, Antje. 2016. *Impact of the type of referring expression on the acquisition of word order variation*. Potsdam: Universitätsverlag Potsdam.

Speyer, Augustin. 2007. Die Bedeutung der Centering Theory für Fragen der Vorfeldbesetzung im Deutschen. *Zeitschrift für Sprachwissenschaft* 26(1): 83–115.

Speyer, Augustin. 2009. Das Vorfeldranking und das Vorfeld-es. *Linguistische Berichte* 219: 323–353.

Speyer, Augustin. 2010. Filling the German vorfeld in written and spoken discourse. In Sanna-Kaisa Tanskanen, Marja-Liisa Helasvuo and Marjut Johansson (eds.), *Discourses in Interaction*, 263–290. Amsterdam: John Benjamins Publishing.

Verhoeven, Elisabeth. 2015. Thematic asymmetries do matter! A corpus study of German word order. *Journal of Germanic Linguistics* 27(01): 45–104.

Ward, Gregory L. and Ellen F. Prince. 1991. On the topicalization of indefinite NPs. *Journal of Pragmatics* 16(2): 167–177.

Weskott, Thomas, Robin Hörnig, Gisbert Fanselow and Reinhold Kliegl. 2011. Contextual licensing of marked OVS word order in German. *Linguistische Berichte* 225: 3–18.

Franziska Münzberg, Sandra Hansen-Morath

Die Wucht und Strömung war immens – wie stark ist der Ellipseneffekt?

Abstract Our corpus study is concerned with subject-verb agreement in contemporary German, more precisely the variation in verb number. We focus on subjects consisting of noun phrases coordinated by the conjunction *und* ('and'). In our samples, both nouns are in singular. Number resolution – i.e., plural verb despite of the singular nouns – can be regarded as the default choice in contemporary German. However, our data show that eliding the second determiner in the subject enhances the probability of using the singular verb. This ellipsis effect is highly significant in German and Austrian texts. It seems to be weaker in Swiss texts. Regression analyses reveal that the ellipsis effect is stronger than both the highly significant influence of subject individuation and the significant effect of subject agentivity.

Keywords Kongruenz, Subjekt, koordiniert, Ellipse, Numerus, Regressionsanalyse

1 Ausgangsfrage und Methode

Bei der Neubearbeitung des „Zweifelsfälledudens" (Duden – Zweifelsfälle 2016) ist die Frage aufgekommen, welche Faktoren besonders stark auf die Kongruenz (oder Korrespondenz) im Numerus zwischen koordiniertem Subjekt und finitem Verb einwirken. Aus einem Teil der Recherchen für den „Zweifelsfälleduden" hat sich in der Folge eine Studie für das IDS-Projekt Korpusgrammatik (http://www1.ids-mannheim.de/gra/projekte/korpusgrammatik.html) entwickelt.[1] Im Fokus stehen Sätze wie dieser:

1 Vielen Dank an Kathrin Kunkel-Razum, Mathilde Hennig und ihr ganzes Team für die gute Zusammenarbeit und für die Möglichkeit, das Dudenkorpus zu verwenden. Für wertvolle Anregungen und Literaturhinweise danken wir unseren Kolleginnen und Kollegen in der Abteilung Grammatik des IDS, besonders Marek Konopka, sowie Klaus Mackowiak, Peter Gallmann, Antje Dammel, Svetlana Petrova und Damaris

(1) Die Wucht und Strömung war immens, die Pulosans wurden meilenweit ins offene Meer getrieben.
(Süddeutsche Zeitung, 19.12.2011: 10, „Verheerende Fluten“)

Formal fallen zwei Besonderheiten auf: Erstens gilt der Artikel *Die* des Subjekts für beide Subjektteile, *Wucht* und *Strömung*. Da vor *Strömung* ein zweiter Artikel mit derselben Form und denselben grammatischen Merkmalen hinzugedacht werden kann, sprechen wir bei diesem Muster mit nur einem Artikel von einer Artikelellipse, einer Ausprägung der Koordinationsellipse.[2] Die zweite formale Besonderheit ist, dass das finite Verb *war* im Singular steht. Wir wollen nachweisen: Dass beides, Ellipse und Singular, zusammen auftritt, ist kein Zufall. Betrachtet man den Numerus des finiten Verbs als abhängige Variable (als „Zweifelsfall“), dann stellt sich dieser Zusammenhang als Auswirkung der Artikelellipse auf den Numerus des finiten Verbs dar. Diesen Effekt nennen wir den Ellipseneffekt.

Vier theoretisch mögliche Muster werden miteinander verglichen:

— Muster 1: Die Wucht und *die* Strömung *war* immens.
— Muster 2: Die Wucht und *die* Strömung *waren* immens.
— Muster 3: Die Wucht und Strömung *war* immens.
— Muster 4: Die Wucht und Strömung *waren* immens.

Unsere Hypothese ist, dass Sätze nach den Mustern 2 und 3 signifikant häufiger vorkommen als Sätze nach den Mustern 1 und 4.

Zur Überprüfung dieser Hypothese dienen eine Voruntersuchung, die die Bedingungen für die Numerusvariation beim finiten Verb klären soll, sowie zwei Hauptuntersuchungen. Die erste der beiden Hauptuntersuchungen soll den Ellipseneffekt überhaupt in einem der Öffentlichkeit zugänglichen Korpus

Nübling. Unser besonderer Dank gilt unserem Kollegen Roman Schneider, der die Datenextraktion für die zweite Hauptuntersuchung vorgenommen hat.

2 Die Einordnung der Konstruktion mit einem Artikel und zwei koordinierten Nomen als Koordinationsellipse ist etabliert, vgl. etwa Hennig (2015: 59), Dammel (2015: 315–317). Wir rechnen also damit, dass die beschriebene Struktur unter diesem Stichwort gesucht wird. Wie allerdings Gallmann in der Dudengrammatik (Duden – Die Grammatik 2016: Rdnr. 1418) – ebenfalls unter dem Stichwort Koordinationsellipse – zeigt, sind NPs mit einem Artikel einerseits und ansonsten identische NPs mit zwei gleichen Artikeln andererseits nicht unbedingt semantisch gleichwertig; vgl. 2.2 (4). Der semantische Unterschied wäre ein Argument für eine Analyse ohne die Annahme einer Ellipse, z. B. als DP. Dass es einen Zusammenhang zwischen Artikelgebrauch und Numeruswahl gibt, kann jedenfalls ganz unabhängig vom zugrundegelegten Grammatikmodell beobachtet werden.

nachweisen. Die zweite Hauptuntersuchung soll Metadaten einbeziehen sowie Aufschluss über weitere wirksame Faktoren geben und zeigen, wie stark die Artikelellipse im Vergleich dazu wirkt.

Nicht im Zentrum der Untersuchung stehen Subjekte ganz ohne Artikel wie in *Wucht und Strömung waren/war immens* (keine Ellipse, sondern „freier Gebrauch ohne Artikel" bei koordinierten NPs: Duden – Die Grammatik 2016: Rdnr. 391).

2 Koordinierte NPs als Subjekte und die Kongruenz im Numerus mit dem finiten Verb – Überblick und Voruntersuchung im Dudenkorpus

2.1 Plural als Normalfall

Die Hypothese, dass das Muster 3 mit Ellipse und Singular häufiger vorkommt als das Muster 4 mit Ellipse und Plural, mag zunächst überraschend klingen. Denn man ist sich einig, dass der Plural bei einem Subjekt mit Koordination im Gegenwartsdeutschen der Normalfall ist *(number resolution)*: Corbett (2000: 198); IDS-Grammatik = Zifonun et al. (1997: Bd. 3, 2388); Donalies (2011); Gallmann in Duden – Die Grammatik (2016: Rdnr. 1602 [Kongruenzregel II]–1613); Grundriss = Eisenberg (2013: Bd. 2, 423, 470); Engel (1996: 188); Helbig/d Buscha (1993: 29); Hoffmann (2016: 452), Mackowiak (2008: 47–50); Duden – Zweifelsfälle (2012: Kongruenz 1.3.1); am vorsichtigsten Duden – Zweifelsfälle (2016: 560 = Kongruenz 1.3.1).

2.2 Bedingungen für die Variation Singular – Plural (abhängige Variable)

Je nachdem, wie viele Seiten die Grammatiker dem Thema Numeruskongruenz bei koordinierten NPs als Subjekten widmen, nennen sie dann aber auch verschiedene Bedingungen dafür, dass beim finiten Verb auch der Singular auftreten kann. Diese Bedingungen gelten für den hier untersuchten Fall, dass beide Subjektteile im Singular stehen. Im Folgenden sind nur solche Bedingungen aufgelistet, die für die Koordination zweier unterschiedlicher Nomen mit der Konjunktion *und* angeführt werden; Zitatsubstantivierungen *(„Jim Knopf und die Wilde 13")* sind nicht berücksichtigt. Die Liste dient dazu, den Ellipseneffekt mithilfe von Korpusanalysen schrittweise von anderen wirksamen Faktoren zu isolieren. So können einige der folgenden Bedingungen bereits durch Suchanfragen ausgeschlossen werden, andere können in den 4 verglichenen Mustern konstant

gehalten werden. Die restlichen Bedingungen werden bei der Belegannotation einbezogen.

1. Wortstellung, Zeitraum, Domäne: Abfolge Verb – Subjekt mit beiden Subjektteilen im Mittelfeld; nach der Dudengrammatik „zuweilen" Singular; in der Standardsprache werde der Plural vorgezogen (Duden – Die Grammatik 2016: Rdnr. 1606); nach der IDS-Grammatik (Zifonun et al. 1997: Bd. 3: 2388) „oft" Singular – der Singular sei hier ein „Normverstoß", „vermutlich als Ergebnis eines Planungsproblems"; nach der 7. Auflage des „Zweifelsfälledudens" ist „der Singular des Verbs möglich, wenn auch seltener als der Plural" (Duden – Zweifelsfälle 2012: Kongruenz 1.3.1); nach Behaghel (1928: 15 = § 808) und Dammel (2015: 307) ist der Faktor Wortstellung eher auf früheren Sprachstufen wirksam, davon im Neuhochdeutschen nur noch Spuren bei Abstrakta in fiktionalen Texten, vgl. Findreng (1976: 209): „plur. Verb bei vorangestelltem Verb fast doppelt so häufig in der Gebrauchssprache (…) wie in der Sprache der schönen Literatur (80 % gegenüber 42 %), während bei Nachstellung die Unterschiede nur gering sind (83 % gegenüber 77 %)".
2. Bedeutung bzw. Referenz: Singular, wenn ein Subjektteil „den anderen Subjektteil inhaltlich einschließt" wie in *er und alle Welt*: Duden – Die Grammatik (2016: Rdnr. 1608), Duden – Zweifelsfälle (2012: Kongruenz 1.3.2); vgl. Schrodt (2005: 243) zur „Termqualität".
3. Referenz allgemeiner: Vorliegen eines einzigen Terms; Probe: Das koordinierte Subjekt ist ersetzbar „durch ein referenzidentisches singularisches Pronomen" (Schrodt 2005: 242); „Synesis" (Helbig/ Buscha 1993: 29).
4. Distributive Lesart: „Man kann auch von einer elliptischen Reihung von zwei Sätzen ausgehen", es liegt also nicht ein einziges koordiniertes Subjekt vor, sondern zwei Subjekte gehören zu zwei Teilsätzen (Duden – Zweifelsfälle 2016: Kongruenz 1.3.1; Hennig 2015: 63–66); vgl. die Wortstellung in *„Die Fachbereichsleiterin zog (…) mit und der neue Kursleiter"* (Mackowiak 2008: 48).
5. (Artikellosigkeit bei) „formelhaften" Subjekten (Behaghel 1928: 18 = § 808, Duden – Die Grammatik 2016: Rdnr. 1609, Mackowiak 2008: 48); wir vermuten, dass dies auch für Formeln mit Artikelellipse wie *die Art und Weise* gilt, vgl. das einschränkende „oft" in „die oft aus Teilen ohne Artikel o. Ä. bestehen" (Duden – Zweifelsfälle 2012: Kongruenz 1.3.3).
6. Artikellosigkeit bei gleichem Genus (Zifonun et al. 1997: Bd. 3: 2388); ähnlich Hoffmann (2016: 452) – zu unterscheiden von der Ellipse eines Artikels. Eine Recherche im Archiv TAGGED-C2 des DeReKo nach den beiden Mustern *Wucht und Strömung werden/wird* lässt darauf schließen, dass dieser Faktor im Gegenwartsdeutschen nicht sehr stark ist: Der Plural des finiten Verbs überwiegt besonders bei konkreten, aber auch bei abstrakten

artikellosen Subjekten mit gleichem Genus deutlich. Auch dieser Recherche zufolge übt das Genus einen Einfluss aus (noch häufiger Plural bei unterschiedlichem Genus).

7. Hoher Abstraktionsgrad der beiden Subjektteile: Singular bei finiten Nebensätzen und nicht substantivierten Infinitiven > substantivierten Infinitiven > „gewöhnlichen Abstrakta" (Duden – Die Grammatik 2016: Rdnr. 1610), vgl. Corbett (2000: 201); Näheres zu Substantivierungen auf *-ung* bei Mackowiak (2008: 47–48); Näheres zur diachronen Entwicklungsrichtung bei Dammel (2015: 314).
8. Niedriger Agentivitätsgrad (Dammel 2015: 308).
9. Subjekt mit Apposition im Singular (Mackowiak 2008: 48, Duden – Zweifelsfälle 2012: Kongruenz 1.3.5: *„Schmidt und Co., Buchdruckerei"* / *„Turm und Brücke – das Hoechster Firmenzeichen –"*).
10. Verbindung mit den Indefinita *kein, jeder, mancher* (Duden – Die Grammatik 2006: Rdnr. 1612, Mackowiak 2008: 48).
11. Bindestrichellipse der Form *Schall- und Wärmedämmung* als Subjekt (Mackowiak 2008: 48, Duden – Die Grammatik 2012: Kongruenz 1.3.4).
12. Gemeinsames Attribut (Duden – Die Grammatik 2012: Kongruenz 1.3.4) zu beiden Subjektteilen, also etwa attributives Adjektiv, Genitiv- oder Präpositionalattribut (Mackowiak 2008: 47, Duden – Die Grammatik 2012: Kongruenz 1.3.4; nach Findreng 1976: 198 nur 16 % Plural, wenn „nur das erste Einzelsubjekt attributive Wörter" hat, allerdings rechnet Findreng 1976: 188–189 wie Behaghel 1928: 17 = § 808 Artikelellipsen mit ein; genauso Duden – Die Grammatik 2016: Rdnr. 1611, einschlägig ist hier nur das letzte Beispiel: *Alle Zerstörungswut und Herrschsucht in uns durfte sich entfalten* [P. Weiss]).
13. Gemeinsames Artikelwort (Duden – Die Grammatik 2012: Kongruenz 1.3.4, Findreng 1976: 188–189, Behaghel 1928: 17 = § 808), was auch wie hier als Artikelellipse interpretiert werden kann (Mackowiak 2008: 47; dort bereits als besonders starker Faktor hervorgehoben).

Die Bedingungen 2 und 3 einerseits und Bedingung 4 andererseits schließen sich gegenseitig aus. Für das Gegenwartsdeutsche lässt sich dieser Widerspruch durch einen Blick in die Dudengrammatik (Duden – Die Grammatik 2016: Rdnr. 1602) recht gut auflösen:

> Kongruenzregel II für Subjekte mit gereihten Subjektteilen:
> (a) Die Reihung gilt gesamthaft als Plural, das finite Verb steht daher ebenfalls im Plural. (b) Die 1. Person rangiert vor der 2. Person, und die 2. Person rangiert vor der 3. Person.
> […]

> Kongruenzregel III: Bei zusammengezogenen Sätzen mit eingesparten finiten Verbformen zählt nur das Subjekt der ausformulierten finiten Verbform.

Bedingung 4 gehört eigentlich nicht in die Liste der Ausnahmen zu Gallmanns Kongruenzregel II (Duden – Die Grammatik 2016: Rdnr. 1605–1612), sondern sie fällt unter Gallmanns Kongruenzregel III für zusammengezogene Sätze. Gallmann und Duden – Zweifelsfälle (2016: 560 = Kongruenz 1.3.1) stimmen darin überein, dass „[d]ie beiden Konstruktionen (…) sich nicht immer eindeutig unterscheiden [lassen]" (Duden – Die Grammatik 2016: Rdnr. 1602). Die vorliegende Untersuchung beschränkt sich auf Fälle, in denen die Kongruenzregel III und damit Bedingung 4 wegen der gewählten Konjunktion und wegen der Wortstellung (Adjazenz der beiden Subjektteile im Vorfeld) u. E. kaum greifen kann.

Für Bedingung 3 formulieren wir die Schrodt'sche Ersatzprobe (Schrodt 2005: 242) um: Die Möglichkeit, das koordinierte Subjekt durch ein referenzidentisches singularisches Pronomen zu ersetzen oder wiederaufzunehmen, ist eine notwendige, aber keine hinreichende Bedingung dafür, dass man bei der Beleginterpretation von einem einzigen Term ausgehen kann. Dass die Bedingung nicht hinreicht, wird sichtbar daran, dass sich das Pronomen *das* mit dem Pronomen *beides* kombinieren lässt – und *beides* setzt die Referenz auf zwei unterschiedliche Entitäten voraus:

(2) Lärmschutz und wirtschaftliche Sicherheit, das sind beides Interessen der Bürger. (Rhein-Zeitung, 24.03.2012: 3)

(3) Ich werfe jetzt mal die Obstplantage und CSI raus, das hat beides in meinen Augen keine überörtliche Bedeutung. (http://de.wikipedia.org/wiki/Diskussion:Wennigsen_(Deister): Wikipedia, 2011)

Dass in (2) der Singular des finiten Verbs keine Option ist, liegt am Prädikativ im Plural.

Die zwei CQP-Abfragen für die Voruntersuchung im Dudenkorpus – eine Abfrage für die beiden Muster mit Ellipse, eine für die beiden Muster ohne Ellipse – haben die folgende Form:

```
[word = "Die"] [c = "noun" & num matches "sg" & word != ".*-"]
[word = "und"] [c = "noun" & num matches "sg" & word != "-.*"]
[vform = "fiv"] within s;
```

Gesucht wurde also das großgeschriebene Wort *Die,* gefolgt von einem Nomen im Singular ohne Bindestrich (Ausschluss von Bedingung 11), gefolgt von *und,* gefolgt von einem Nomen im Singular ohne Bindestrich, gefolgt von einer finiten Verbform. Bei der Suche für die Muster ohne Ellipse folgte auf *und* das Wort *die.* Gefunden wurden in ca. 4 Milliarden morphosyntaktisch annotierten Wortformen 4.382 Treffer für die Muster ohne Ellipse und 5.701 Treffer für die Muster mit Ellipse. Die Treffer waren jeweils nach dem Zufallsprinzip angeordnet, und jeweils die ersten paar hundert wurden durchgesehen, um explorativ einen ersten Überblick über die Trefferqualität zu erhalten.

Heraus kam zunächst, dass der Numerus in einigen Belegen nicht anders hätte gewählt werden können. Um interessante Aussagen über Numerusvariation zu machen, muss man wohl die folgenden Fälle ausschließen:

— Mit dem Subjekt wird klar auf ein und dieselbe Entität (Extremfall: Einzelperson) referiert – und sowohl der Singular des finiten Verbs als auch die Artikelellipse markieren das (Duden – Die Grammatik 2016: Rdnr. 1418, Zifonun et al. 1997: Bd. 3, 2388); der Ersatz durch ein Pronomen im Plural ist unmöglich:

> (4) Die Sportwissenschaftlerin und Sportmedizinerin erklärt, dass bei solchen Extrembelastungen der Druck auf die Gefäße viel zu groß sei. (Mannheimer Morgen, 22.03.2004, o. S., „Statt Stress in der Muckibude lieber langsam laufen“)

— Das Prädikativ zu einem Kopulaverb steht im Plural (Duden – Die Grammatik 2016: Rdnr. 1632); vgl. (2).
— Mit dem Subjekt wird klar auf zwei Entitäten (Extremfälle: Einzelpersonen, mit einem geografischen Eigennamen Benanntes; vgl. Behaghel 1928: 18 = § 808) referiert; der Ersatz durch ein Pronomen im Singular ist unmöglich und das finite Verb wird in den Plural gesetzt:

> (5) Die ÖBB und die Gendarmerie baten um sachdienliche Hinweise der Bevölkerung. (Der Standard, 06.05.2005: 8, „Sechsmal mehr Lawinentote“)

— Das Verb ist reziprok und steht im Plural:

> (6) Die Impfung und die Krebsfrüherkennung ergänzen sich und gewährleisten so die bestmögliche Vorsorge vor Gebärmutterhalskrebs. (news aktuell = dpa-Tochter, 29.03.2007)

— Ohne Reflexivpronomen, aber wohl vergleichbar: *auseinanderliegen, übereinstimmen* ohne PP *mit …* als Ergänzung. Letzteres dürfte nach Schrodt (2005: 239) noch eher den Singular zulassen, aber auch *übereinstimmen* ohne *mit* setzt in all seinen Lesarten die Referenz auf zwei Entitäten voraus. Ein Beleg für die Lesart ‚sich einig sein':

(7) Die Bundesregierung und die Strombranche stimmen darin überein, dass angesichts stark verringerter Abfallvolumina ein einziges Endlager in Zukunft genügt.
(Handelsblatt, 09.02.2000: 2, „ABGESCHNITTEN: […]")

Ein wenig überraschendes Ergebnis der Voruntersuchung im Dudenkorpus ist, dass die Grundregel (2.1) für Sätze mit zwei definiten Artikeln und ohne (weitere) Attribute im Subjekt stimmt: In der weitaus überwiegenden Zahl der ausgezählten Treffer ohne Artikelellipse steht das finite Verb im Plural. Auch wenn man die 127 Fälle unberücksichtigt lässt, in denen klar auf zwei Entitäten referiert wird und der Ersatz des koordinierten Subjekts durch *das* kaum möglich erscheint (Bedingung 3 also nicht wirken kann), bleiben noch 102 Pluralbelege gegenüber 7 Singularbelegen übrig. Die Bedingungen 2, 3, 7 und 8, die sich nicht durch die Suchanfrage ausschließen lassen, treffen also auf wenige Sätze zu (Bedingung 3 trifft zwar oft zu, aber für Bedingung 2 wurde nur ein Beleg gefunden), und/oder sie wirken nicht sehr stark und/oder sie wirken nicht unabhängig vom Ellipseneffekt. Bei den 162 ausgezählten Treffern mit Artikelellipse wiederum (Bedingung 13) wurden Subjekte mit zwingender Referenz auf eine einzige Entität (Ersatz durch ein Pronomen im Plural unmöglich) immerhin 52-mal gefunden. Andere Bedingungen als 2, 3, 7, 8 und 13 konnten wegen des Suchdesigns nicht wirken.

An den Ausschlusskriterien „eindeutige Referenz auf eine Entität / zwei Entitäten" (vgl. (4) und (5)), die besonders deutlich bei Personenbezeichnungen zutage treten, zeigt sich: Von einer Numerusvariation beim Verb zu sprechen, wird problematisch, wenn die Nomen nicht abstrakt genug, zu belebt, zu „individuativ" sind (vgl. Eisenberg: Bd. 2, 140–148, Gunkel et al. 2016: 295). Graduell abgestuft ist das in einer Skala „animacy/individuation" (Dammel 2015: 294) bzw. „hierarchy of individuation": „human > anim > count > mass > abstract > (nominalizations)" (Dammel 2015: 318).

Im Folgenden ist immer vereinfachend vom „Abstraktionsgrad" der beiden Nomen die Rede.

2.3 Bedingungen für die Variation Artikelellipse – keine Artikelellipse (Prädiktorvariable)

Im Gegenwartsdeutschen ist die Hauptrestriktion für Koordinationsellipsen die, dass ausgedrückte und eingesparte Teile in der Regel dieselben grammatischen Merkmale aufweisen. Das ist in der Suchanfrage der Voruntersuchung zum Muster ohne Ellipse bereits berücksichtigt: Gesucht wurde ja nach Nomen im Singular, die beide mit dem definiten Artikel *die* verbunden sind. Die Belege aus der Voruntersuchung zeigen aber, dass im speziellen Fall der Artikelellipse noch weitere Restriktionen zu gelten scheinen. Wenn nicht ohnehin beide Nomen zwingend auf dasselbe Individuum referieren (Referenzidentität bei niedrigem Abstraktionsgrad, 52 ausgeschlossene Belege), so gilt im Singular tendenziell: Je höher der Abstraktionsgrad der Nomen, desto wahrscheinlicher die Artikelellipse (vgl. Heycock und Zamparelli 2005: 211, 214) – mit einem interessanten Ausreißer (*die CDU und SPD* ist im Dudenkorpus im Vergleich zu *die CDU und die SPD* gut belegt). Typisch für Sätze mit Artikelellipse im Subjekt und ohne zwingende Referenzidentität der beiden Nomen ist dieser:

(8) Die Größe und Anordnung wird vom Heraldiker überprüft.
(Freie Presse, 03.03.2015: 13, „Die ersten Vorschläge für das Hartmannsdorfer Wappen“)

Aussortiert werden müssen natürlich die vielen Belege, deren zweites Nomen primär artikellos (Duden – Die Grammatik 2016: Rdnr. 397) ist; hier gibt es keine Artikelellipse:

(9) Die EU und Russland haben die letzten verbleibenden bilateralen Fragen für einen Beitritt Russlands zur Welthandelsorganisation WTO gelöst.
(NZZ, 22.10.2011: 29, „Der WTO-Beitritt Russlands rückt näher“)

Für die Hauptuntersuchungen legen wir zunächst fest, dass alle Vergleichssätze koordinierte Nomen mit hohem Abstraktionsgrad als Subjekt aufweisen sollen. Die Wahrscheinlichkeit dafür erhöhen wir dadurch, dass wir nach koordinierten Nomen suchen, von denen das erste auf *-ung* endet. So wird Bedingung 7, die ja nicht ausgeschlossen werden kann (jedes Nomen hat irgendeinen Abstraktionsgrad und steht irgendwo in der „hierarchy of individuation“), nach Möglichkeit konstant gehalten. Belege mit weniger abstrakten Nomen, die in der „hierarchy of individuation“ weiter links stehen (etwa *Regierung*), sollen markiert und extra ausgewertet werden, damit das Zusammenspiel zwischen Bedingung 7 und dem Ellipseneffekt beobachtet werden kann.

3 Hauptuntersuchung 1 im DeReKo, Archiv TAGGED-C2: der Numerus in Abhängigkeit von der Artikelellipse

3.1 Korpus und Suchanfragen

Das öffentlich zugängliche Archiv TAGGED-C2 des DeReKo (Institut für Deutsche Sprache 2016a) enthält rund 1,4 Milliarden Wortformen; es besteht aus Ausgaben der „VDI nachrichten“ und Pressetexten aus Deutschland, Österreich und der Schweiz aus den Jahren 2010–2014 (Institut für Deutsche Sprache 2016b). Von den 17 Teilkorpora des Archivs TAGGED-C2 überschneiden sich 3 überregionale Tageszeitungen mit den 20 Zeitungskorpora des Dudenkorpus (darüber hinaus enthält das Dudenkorpus auch Sachbücher und fiktionale Texte). Die Texte wurden mit dem Connexor Machinese Phrase Tagger (Connexor Oy 2011–2016) morphosyntaktisch annotiert. Die Suchanfragen lauten

> Die /+w1,so (MORPH(N -PL) /wo *ung) /+w1 „und“ /+w1 MORPH(N -PL) /+w1,so MORPH(V -INF -PCP)
>
> und
>
> Die /+w1,so (MORPH(N -PL) /wo *ung) /+w1 „und“ /+w1 die /+w1 MORPH(N -PL) /+w1,so MORPH(V -INF -PCP)

Wegen des letztlich doch seltenen Auftretens von Bindestrichellipsen wurde die Einschränkung „keine Bindestrichellipse“ in die händische Belegannotation verschoben.

3.2 Belegannotation

Nicht gewertet wurden Sätze mit folgenden Merkmalen (vgl. die Bedingungen für die Numerusvariation 2.2): Bindestrichellipse; formelhaftes Subjekt (*Die Forschung und Lehre, Die Forschung und Entwicklung* – beides nur mit Singular); koordinierte NP ist kein Subjekt; ein primär artikelloses Nomen im Datensatz für Ellipsen (Muster *Die EU und Russland*); kein finites Verb; Subjektteil im Plural; Prädikativ, das den Numerus beeinflusst haben könnte; Subjekt bezeichnet zwei Individuen *(Die Bedienung und die Kundin).* Der in der Voruntersuchung häufige Fall, dass das komplexe Subjekt insgesamt ein Individuum bezeichnete, kam in der Hauptuntersuchung nicht mehr vor. Hier hat sich die Einschränkung auf erste Nomen mit der Endung *-ung* bewährt. Auch Sätze, auf die Bedingung 2 zutraf (ein Subjektteil schließt den anderen inhaltlich ein), wurden nicht mehr

gefunden. Ausgefiltert wurden Traueranzeigen, da diese Textsorte sehr unregelmäßig übers Korpus verteilt war. 322 von 643 Sätzen blieben übrig. Extra ausgezählt wurden 72 Sätze, in denen mindestens ein Subjektteil nicht abstrakt genug war bzw. in der „individuation hierarchy" zu weit links stand, um in die Wertung mit einzugehen (meist sog. *committee nouns*, d. h. Bezeichnungen für Personengruppen: *Die Bevölkerung und die Politik; Die Bundesregierung und die EU-Kommission;* vgl. 2.3). Jeweils das Nomen mit dem niedrigeren Abstraktionsgrad legte den Abstraktionsgrad der gesamten NP fest. Dieser Teil der Annotation ist der subjektivste und daher auch der problematischste. Reziproke Verben kamen nicht vor, aber ein Beleg wurde ausgefiltert, weil für das Subjekt zwei Rollenträger verlangt waren.[3]

3.3 Ergebnisse

Zunächst bestätigt sich die Hypothese aus der Voruntersuchung, dass die Artikelellipse im Singular besonders bei Abstrakta auftritt (Tabelle 1):

Tabelle 1: Artikelellipse im Singular bei hohem Abstraktionsgrad.

n = 322	ohne Ellipse	mit Ellipse	gesamt
geringerer Abstraktionsgrad (meist *committee nouns* vom Typ *Bevölkerung, Regierung*)	61 = 85 %	11 = 15 %	72 = 100 %
hoher Abstraktionsgrad (meist Verbalsubstantive vom Typ *Beratung, Stellenvermittlung*)	65 = 26 %	185 = 74 %	250 = 100 %

Und es bestätigt sich die Hypothese, dass eine Artikelellipse im Subjekt den Singular beim finiten Verb begünstigt (Tabelle 2). In Klammern stehen die Zahlen vor der Ausfilterung nach dem Abstraktionsgrad.

Tabelle 2: Numerus und Ellipse, Hauptuntersuchung 1.

n = 250 (322)	finites Verb im Singular	finites Verb im Plural	gesamt
ohne Ellipse	26 % (14 %)	74 % (86 %)	100 % (100 %)
mit Ellipse	91 % (89 %)	9 % (11 %)	100 % (100 %)

Der Assoziationsplot (vgl. Cohen 1980, Friendly 1992, Meyer et al. 2005) in Abbildung 1 zeigt, dass auch noch nach der Ausfilterung belebter bzw. „zu konkreter"

3 *Die Dichtung und die Liebe gehörten für die junge Frau zusammen* (…).

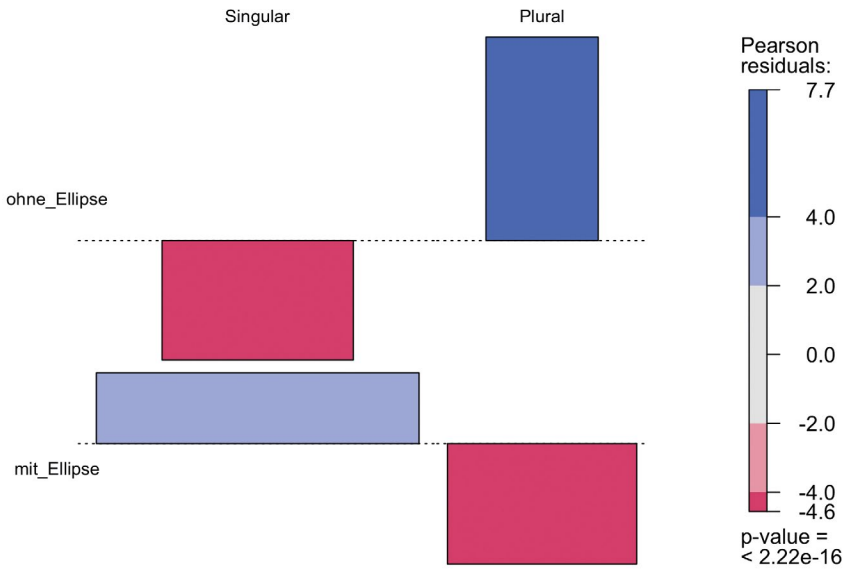


Abbildung 1: Assoziationsplot zur ersten Hauptuntersuchung im Korpus TAGGED-C2: Die Artikelellipse wirkt sich auf den Numerus des finiten Verbs aus.

Subjekte in Sätzen ohne Ellipse der Plural signifikant überrepräsentiert ist, der Singular signifikant unterrepräsentiert. In Sätzen mit Ellipse kehrt sich das Verhältnis um.[4]

4 Hauptuntersuchung 2 im DeReKo, KoGra-Untersuchungskorpus: weitere Einflussfaktoren

In der zweiten Hauptuntersuchung gehen wir der Frage nach, ob neben dem Ellipseneffekt auch die Metadaten Land, Zeit und Domäne einen Einfluss auf den Numerus haben. Außerdem nehmen wir nun die Verben genauer unter die Lupe.

4 Der Assoziationsplot wurde über das statistische Auswertungstool KoGra-R (Institut für Deutsche Sprache 2015; vgl. Hansen-Morath et al. [in Vorbereitung]) erstellt. Der Plot stellt die standardisierten Pearson-Residuen der Häufigkeiten von Singular und Plural in Abhängigkeit vom Vorhandensein einer Ellipse dar. Balken oberhalb der gepunkteten Linie bedeuten, dass die Werte höher sind als erwartet, Balken unterhalb der Linie bedeuten, dass die Werte niedriger sind als erwartet. Die Breite der Balken spiegelt die erwartete Frequenz der Realisierungen wider. Signifikante Pearson-Residuen werden im Plot rot bzw. blau eingefärbt (vgl. ebd.).

4.1 Korpus und Suchanfragen

Das KoGra-Untersuchungskorpus besteht aus knapp 8 Milliarden morphosyntaktisch annotierten (Connexor, TreeTagger; vgl. Schmid 1995) Wortformen aus dem DeReKo (Institut für Deutsche Sprache 2014). Zu den hinterlegten Metadaten gehören Land, Region, Datum und Domäne (Näheres über die Begriffe Region und Domäne, den Korpusaufbau und die Abfragemöglichkeiten in Bubenhofer et al. 2014: 21–117; zur aktuellen Größe und Struktur vgl. das grammis-Modul „Korpusgrammatik" unter https://grammis.ids-mannheim.de/korpusgrammatik).[5] Mit dem Dudenkorpus überschneiden sich 8 von 60 Teilkorpora. Die Texte des Korpus TAGGED-C2 sind in der KoGra-Datenbank enthalten. Bei den Suchanfragen gab es einen Unterschied zu den beiden anderen Untersuchungen: Um von vornherein nach Numerus getrennte Ergebnisse zu erhalten, wurde die Position „finite Verbform" eingeschränkt auf 1. Verbformen, die auf *te/*ten[6] enden, sowie 2. die Verbformen *kann/können* ODER *muss/müssen* ODER *soll/sollen* ODER *darf/dürfen* ODER *hat/haben* ODER *wird/werden*. Das Verb *sein* wurde nicht berücksichtigt, damit es weniger Belege gibt, in denen ein Prädikativ den Numerus beeinflusst. Ansonsten wurde auf der Grundlage der Connexor-Annotation dasselbe gesucht wie in der ersten Hauptuntersuchung (vgl. 3.1).

4.2 Belegannotation

Die Belege wurden zunächst annotiert wie in der ersten Hauptuntersuchung. Beim Abstraktionsgrad haben wir uns aus pragmatischen Gründen wieder für 0 (= geringer Abstraktionsgrad) oder 1 (= hoher Abstraktionsgrad) entschieden. Interessant wäre vielleicht auch eine Abstufung verschiedener Abstraktionsgrade gewesen: 0 für Belebtes, 1 für „gewöhnliche" Abstrakta und 2 für Verbalsubstantive, die nicht nur mit dem Suffix *-ung* gebildet sind, sondern tatsächlich Tätigkeiten oder Vorgänge bezeichnen. Allerdings war nach der ersten Hauptuntersuchung zu erwarten, dass typische Verbalsubstantive auch das Gros der Nomen mit ausreichendem Abstraktionsgrad ausmachen würden. Sätze aus Traueranzeigen wurden nicht ausgefiltert. Reziproke Verben wurden nicht

5 Das KoGra-Untersuchungskorpus stellt eine Auswahl aus DeReKo-Texten dar, die mit den genannten Metadaten angereichert wurden.

6 Schwache Präteritumendungen und nicht etwa Präsensendungen *t/*en, weil bei der Suche nach *en auch Präteritumformen wie *fragten* oder *gaben* gefunden und als Plural gezählt würden, nicht aber bei der Suche nach *t die entsprechenden Singularformen *fragte, gab*. Auch *t ist mehrdeutig (3. Pers. Sg. / 2. Pers. Pl.). Das hätte aufwendige Annotationsarbeit verursacht.

gefunden. 11 Belege mussten ausgefiltert werden, weil das Verb für das Subjekt zwei Rollenträger verlangte.[7] Anders als in der ersten Hauptuntersuchung wurden nun zusätzlich Diathese, Agentivität, Verbtyp (Voll-, Hilfs-, Modalverb, modifizierendes Verb) und bei den Suchen nach Verben auf *te/*ten auch die Übereinstimmung mit dem Suchmuster (3. Pers. Indikativ Präteritum schwacher Verben) berücksichtigt: Endet der Verbstamm auf *-t,* muss der Beleg ausgefiltert werden, denn in Konkurrenz etwa zum Plural *bieten* steht ja neben dem Konjunktiv I *biete* vor allem der Indikativ Präsens *bietet,* und diese Form würde über die Suche nach *te nicht gefunden. Aus demselben Grund wurde auch die Verbform *taten* nicht akzeptiert (Singular zu *taten: tat,* nicht *tate*).

4.3 Ergebnisse

Die Ergebnisse für die beiden Variablen Numerus und Ellipse in Tabelle 3 sehen ähnlich aus wie bei der ersten Hauptuntersuchung, nur dass es insgesamt mehr Pluralbelege gibt:

Tabelle 3: Numerus und Ellipse, Hauptuntersuchung 2.

n = 842	finites Verb im Singular	finites Verb im Plural	gesamt
ohne Ellipse	38 = 16 %	204 = 84 %	242 = 100 %
mit Ellipse	503 = 84 %	97 = 16 %	600 = 100 %

Eine Auswahl an Belegen für jedes der vier Muster:

(10) a. Die Durchführung und die Organisation wird einem privaten Büro übertragen. (St. Galler Tagblatt, 14.05.1999, „Altersheim-Anbau nimmt Gestalt an“)
b. Die Verantwortung und die „Haftung“ muss in den jeweiligen Ländern bleiben. (Potsdamer Neueste Nachrichten, 05.05.2010, „‚Die Lage kann sehr schnell eskalieren‘ […]“)

(11) a. Die Abfertigung und die Gepäckbeförderung werden optimiert (…). (Frankfurter Allgemeine Zeitung, 15.05.1997, „Die ‚Star Alliance‘ geht an den Start“)

7 Etwa *Die Dorferneuerung und die Gemeinde müssen zusammen arbeiten* (…) u. Ä. 8 ausgefilterte Belege entsprachen dem Muster 2, 2 dem Muster 3 und 1 Beleg (bei dem auch unklar ist, ob es sich um eine Ellipse handelt: *Himmelfahrt* artikellos?) dem Muster 4: *Die Auferstehung und Himmelfahrt gehörten ganz zusammen* (…). Eine falsche Ellipse wurde gefunden und mitgezählt: *Die Sanierung und Ausbau soll* (…).

b. Die Bildung und die Erziehung müssen im Vordergrund stehen. (Rhein-Zeitung, 01.02.2014: 22, „Sekundarschule sorgt für heiße Diskussion“)

(12) a. Die Erhaltung und Erweiterung wird vorwiegend durch Samen- und Pflanzenaustausch mit anderen Botanischen Gärten sichergestellt. (Botanischer Garten (Rostock). In: Wikipedia - URL: http://de.wikipedia.org/wiki/Botanischer_Garten_(Rostock): Wikipedia, 2011)
b. Die Beratung und Vermittlung soll im Mai beginnen. (Rhein-Zeitung, 28.03.1998, „Ziel ist: Arbeit statt Sozialhilfe“)

(13) a. Die Hochblätterfärbung und Blütenbildung werden in den Gewächshäusern durch Lichteinwirkung bestimmt. (Schweriner Volkszeitung, 24.12.2009: 13, „Weihnachtsstern steht“)
b. Die Entwicklung und Umsetzung sollen mindestens 3,5 Mrd. kosten (…) (Nürnberger Nachrichten, 03.12.2005, „Fertig zum Start? Bei Galileo zögern Mittelständler“)

Zunächst ist zu klären, welchen Einfluss die Metadaten auf die Variation zwischen Singular und Plural haben: das Land (belastbare Zahlen haben wir zu Deutschland, Österreich und der Schweiz; zu wenige Belege für Luxemburg), das Jahrzehnt (belastbare Zahlen zu den 1990ern, 2000ern und 2010ern, nur Einzelergebnisse für die 1960er und 1980er) und die inhaltliche Domäne (Fiktion, Kultur, Mensch, Politik, Technik). Die statistischen Analysen ergeben, dass der Ellipseneffekt in den Schweizer Daten schwächer ist.[8] Ansonsten gibt es keine signifikanten Einflüsse.

Diese erstaunliche Aussage müssen wir ein bisschen einschränken: Sie bezieht sich eben auf unseren Datensatz. Erstens dokumentiert er keine Entwicklungen über die Jahrhunderte hinweg. Sowohl die Bedingungen für Ellipsen überhaupt (Hennig 2010) als auch die für die Numerusvariation (Dammel 2015) haben sich über verschiedene Sprachstufen entscheidend verändert. Da *number resolution* auf früheren Sprachstufen weniger formalisiert ist, ist etwa für das Frühneuhochdeutsche mit einem weniger spektakulären Ellipseneffekt zu rechnen (vgl. Dammel 2015: 315–317). Nach Behaghel (1928: 17= § 808 A I 2 α) ist der Ellipseneffekt allerdings nicht aufs Neuhochdeutsche beschränkt.

Zweitens enthält unser Datensatz nur einen einzigen Beleg aus der Domäne Fiktion, der schon wegen der Verbform *taten* aussortiert werden musste, vgl. 4.2:

8 Rechnerisch überprüft anhand der Pearson-Residuen und visuell mit einem Assoziationsplot.

Die Bewegung und die Waldluft taten ihm gut. Sätze in fiktionalen Texten beginnen eben nicht typischerweise mit zwei koordinierten Verbalabstrakta.

Für die Numerusvariation ist es irrelevant, ob ein Verb dem Suchmuster *te/*ten entstammt oder der Suche nach den frequenten Verben *können, müssen, sollen, dürfen, haben* und *werden.* Auch ob das finite Verb ein Vollverb, ein Modalverb, ein modifizierendes Verb, ein Perfekt- oder ein Passivhilfsverb ist, hat keinen Einfluss auf die Wahl des Numerus.[9]

Der folgende Assoziationsplot (Abbildung 2) zeigt: Genauso wenig wichtig ist, ob es sich um einen Passivsatz handelt (Agentivitätsgrad 0). Bei den Sätzen im Aktiv hingegen macht es einen Unterschied, welchen Grad an Agentivität das Verb von seinem Subjekt verlangt. Bei Verben bzw. Konstruktionen, deren Subjektaktant die Rolle eines Auslösers oder gar Verursachers hat, wie *dazu führen, vor Herausforderungen stellen, Spaß machen, Probleme bereiten, fördern, zu schaffen machen, sorgen für* u. ä. (Agentivitätsgrad 2), ist der Ellipseneffekt schwächer als bei Verben wie *stimmen, kosten, dauern, erfolgen, stattfinden, sich verzögern, sein* (Agentivitätsgrad 1): Beim Agentivitätsgrad 2 gibt es in den Sätzen mit Ellipse 41 Singularbelege gegenüber 23 Pluralbelegen. Ansonsten überwiegt der Singular sehr viel stärker, sodass die tatsächliche Häufigkeit des Singulars beim Agentivitätsgrad 2 niedriger ist als der statistisch erwartete Wert.[10]

Die referierten Zahlen beziehen sich alle auf Belegsätze, deren Subjekten bei der Belegannotation der Abstraktionsgrad 1 zugeordnet worden ist; dieser Faktor wurde also konstant gehalten. In einem weiteren Schritt sollten jedoch statistische Modelle berechnet werden, die den Abstraktionsgrad der Nomen als möglichen Einflussfaktor im Zusammenspiel mit der Artikelellipse berücksichtigen. Dazu müssen nun wieder diejenigen Belege in den Blick genommen werden, die wegen eines zu niedrigen Abstraktionsgrades (0) zunächst nicht berücksichtigt worden sind. Der Abstraktionsgrad 0 wurde in der überwiegenden Mehrzahl der Fälle vergeben, weil es sich um *committee nouns* handelte; seltener kamen typische Konkreta wie *Lenkung und Hinterachse* vor. Nur sehr vereinzelt gab es Zweifel, ob nicht doch zwingend auf dieselbe Entität referiert wird (*die Geburtsabteilung und Gynäkologie, die Ausstellung und Börse; die Stimmung und Atmosphäre* mit Singular des finiten Verbs, aber auch *die Stimmung und die Atmosphäre* mit Plural des finiten Verbs).

9 Rechnerisch überprüft anhand der Pearson-Residuen und visuell mit Assoziationsplots.

10 Dieser Effekt wurde außerdem mithilfe einer Regressionsanalyse bestätigt, deren Ergebnisse hier aus Platzgründen nicht detailliert vorgestellt werden können. Zusammengefasst ergibt die Analyse, dass die Wahrscheinlichkeit für den Singular bei elliptischen Konstruktionen höchstsignifikant ansteigt und bei Verben, die einen Agentivitätsgrad von 2 aufweisen, signifikant fällt.

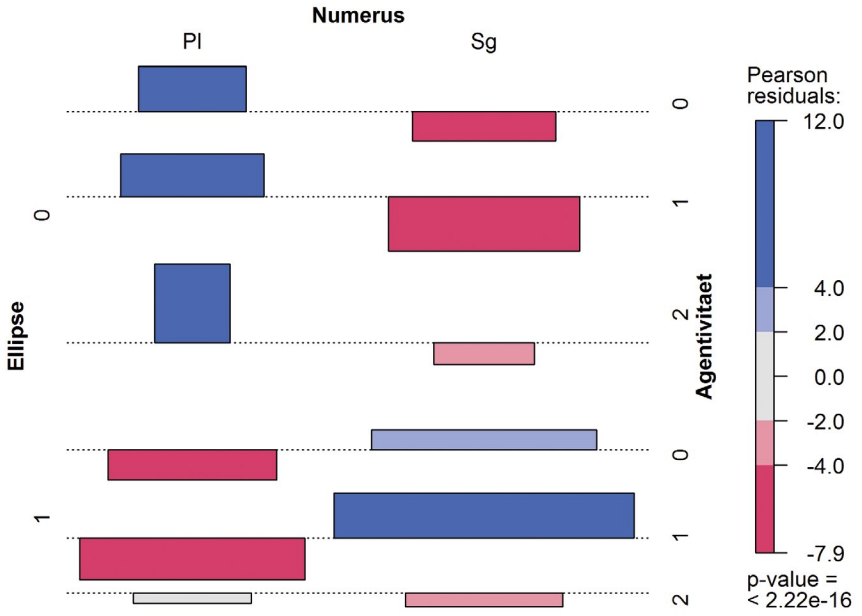


Abbildung 2: Assoziationsplot zum Ellipseneffekt bei unterschiedlichen Graden der Agentivität: stärkerer Ellipseneffekt bei Passiv (Agentivitaet 0) und bei wenig agentivischen Subjekten (Agentivitaet 1); schwächerer Ellipseneffekt bei stärker agentivischen Subjekten (Agentivitaet 2).

Um zu überprüfen, ob neben dem Vorhandensein einer Ellipse der Abstraktionsgrad einen Einfluss auf die Variation des Numerus hat, wurden mehrere logistische Regressionsmodelle berechnet (vgl. Dobson 1990, Hastie und Pregibon 1992, McCullagh und Nelder 1989). Hierzu wurde der Datensatz in zwei gleich große Teile geteilt.[11] Der erste Teil diente der Entwicklung des „besten" Modells (durch Einschluss bzw. Ausschluss bestimmter Einflussfaktoren = Prädiktoren). Der zweite Teil diente der Evaluation des ausgewählten Modells. Die Modellselektion auf dem ersten Teil des Datensatzes ergab, dass das Modell mit beiden Faktoren (Ellipse und Abstraktionsgrad) und ohne Aufnahme der Interaktion zwischen beiden Faktoren das beste Modell ist.[12] Die Modelle sagen für jeden Fall eine

11 Die Zuordnung der Fälle geschah zufällig.

12 Das Modell mit Interaktion weist dieselben Pseudo-R-Quadrate auf (= Anteile des durch die Prädiktoren aufgeklärten Informations- bzw. Variationsanteils), während das „einfachste" Modell mit nur einem Faktor (Ellipse) niedrigere Pseudo-R-Quadrate hat. Die Werte sind so zu interpretieren, dass ein Modell mit größeren Indizes einen besseren Fit gegenüber einem anderen Modell mit geringeren Werten aufweist: Das ausgewählte Modell mit beiden Prädiktoren ohne Aufnahme der Interaktion ergibt einen McFadden-Index von 0,53. Bei dem Modell mit zwei Prädiktoren und Interakti-

Wahrscheinlichkeit für die Ausprägung des Numerus vorher.[13] Die vorhergesagten Fälle werden mit den tatsächlichen Beobachtungen verglichen und der Anteil an korrekt klassifizierten Fällen berechnet. In den Modellen mit beiden Prädiktoren (Ellipse und Abstraktionsgrad) mit und ohne Interaktion werden 89 Prozent der Daten korrekt vorhergesagt. Interessanterweise liegt die Vorhersagekorrektheit in dem Modell mit einem Prädiktor (Ellipse) ebenfalls bei 89 Prozent. Aufgrund der Ergebnisse aus den Modellvergleichen wird auf dem zweiten Teil der Daten das Modell mit beiden Prädiktoren ohne Interaktion berechnet. Die Kennwerte der logistischen Regression dieser Berechnung lauten wie folgt (Tabelle 4):

Tabelle 4: Statistische Kennwerte des logistischen Regressionsmodells für die Analyse der Variation des Numerus in Abhängigkeit von den Faktoren Ellipse (Ellipse1 = Ellipse liegt vor) und Abstraktionsgrad (Abstraktionsgrad1 = hoher Abstraktionsgrad). Signifikanz (p < …): '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1.

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	Signif.
(Intercept)	-3.9180	0.3450	-11.355	< 2e-16	***
Ellipse1	3.5943	0.2890	12.438	< 2e-16	***
Abstraktionsgrad1	1.9569	0.3504	5.585	2.33e-08	***

Sowohl das Vorhandensein einer Ellipse als auch der Faktor Abstraktionsgrad wirken höchstsignifikant (p < 0,001) auf den Numerus des finiten Verbs ein (= mehr Singular). Der McFadden-Index liegt in diesem Modell bei 0,54, der Nagelkerke-Index bei 0,7. Durch das Modell werden insgesamt 88 Prozent der Fälle korrekt vorhergesagt. Die Werte der Estimates bestätigen, dass der Ellipseneffekt stärker wirkt als der Effekt durch den Abstraktionsgrad.

onsaufnahme liegt der Wert ebenfalls bei 0,53. Das Modell, in dem lediglich der Faktor Ellipse untersucht wurde, weist einen McFadden-Index von 0,49 auf. Der Nagelkerke-Index liegt bei dem Modell mit zwei Faktoren ohne Interaktion bei 0,69, bei dem Modell mit Interaktion und zwei Prädiktoren ebenfalls bei 0,69. Das Modell mit einem Faktor (Ellipse) weist einen Nagelkerke-Index von 0,66 auf. Mithilfe von ANOVAs werden die berechneten Modelle miteinander verglichen. Die Vergleiche zeigen, dass das Modell mit beiden Faktoren (Ellipse und Abstraktionsgrad) ohne die Aufnahme der Interaktion die Daten signifikant besser erklärt als das komplexe Modell mit Aufnahme der Interaktion und als das „einfachste" Modell mit einem Hauptfaktor (Ellipse).

13 Als Schwellenwert dient hier üblicherweise eine vorhergesagte Wahrscheinlichkeit von 0,5. Ist die vorhergesagte Wahrscheinlichkeit höher als der Schwellenwert, gehen wir davon aus, dass das Modell eine Singularform vorhersagt.

5 Offene Liste offener Punkte

Bei (mindestens) 13 Faktoren (vgl. 2.2), die nach dem derzeitigen Forschungsstand die Ausprägung des Numerus mitbestimmen können, wäre es natürlich viel zu gewagt, wenn man den Faktor Ellipse gleich als den stärksten davon bezeichnen wollte. Um es mit Hennigs Worten zu sagen: Hier gibt es auch innerhalb von Standardvarietäten „konfligierende Teilsysteme" (Hennig 2017: 34, 42–43) und sich gegenseitig verstärkende Faktoren. Ungeklärt bleibt insbesondere, wie weit unser Begriff „Ellipseneffekt" gefasst werden sollte. Untersucht haben wir ja nur definite Artikel (und das auch nur bei Feminina). Ein rascher Blick auf Ellipsen in NPs mit anderen Wortformen, die im Dudenkorpus als Determiner analysiert sind, zeigt bei der Frage nach der Kongruenz mit dem finiten Verb keine Unterschiede zwischen definiten Artikeln, Demonstrativa und Possessiva, aber hier haben wir keine genauen Zahlen gesammelt. Wir wissen nicht, wie stark der Einfluss der Koordinationsellipse von Pronominaladjektiven und von gewöhnlichen Adjektiven, soweit diese im Singular in Subjekten artikellos vorkommen, im Vergleich zur Artikelellipse ist; dazu haben wir keine eigenen Zahlen. Wir halten es für denkbar und praktisch, all das unter „Ellipseneffekt" zusammenzufassen, sollten sich dabei ähnliche Zahlenverhältnisse ergeben. Von unseren koordinierten NPs mit zwei Nomen zu trennen sind allerdings Konstruktionen wie *die technische und künstlerische Begabung* (Duden – Zweifelsfälle 2012: Kongruenz 1.3.4). Allein die Frage, ob überhaupt eine Ellipse vorliegt, ist in solchen Konstruktionen mit nur einem Nomen viel schwieriger.

Auch bleibt zu untersuchen, wie stark der Ellipseneffekt inzwischen bei Subjekten im Mittelfeld ist – in Verbletztsätzen und natürlich besonders in Verbzweitsätzen (2.2, Bedingung 1).

Schließlich ist es sowohl im statistischen Sinne als auch im Rahmen der linguistischen Analyse möglich, abweichend von der traditionellen Formulierung des Zweifelsfalls („Singular oder Plural?") gar nicht den Numerus des finiten Verbs, sondern die Ellipse im koordinierten Subjekt als abhängige Variable zu betrachten: Je nach der Intention der Schreibenden lassen sich Sätze mit einem Verb im Singular manchmal durch eine Artikelellipse unauffälliger machen und Sätze mit einem Verb im Plural durch das Hinzusetzen eines zweiten Artikels.

6 Zusammenfassung

Unter den vier Mustern

(14) Die Aufregung und die Spannung ist gross. (St. Galler Tagblatt, 12.06.2010: 47; „Das schaff' ich – oder eben doch nicht?")
(15) Die Erwartung und die Aufgabe waren klar: Dieses Spiel musste gewonnen werden. (St. Galler Tagblatt, 08.02.2010: 38; Bütschwiler bezwingen Appenzeller)
(16) Die Einteilung und Farbgebung kann sich dem Inhalt anpassen oder umgekehrt. (St. Galler Tagblatt, 26.11.2012: 42; Raum – verschieden umgesetzt)
(17) Die Erstellung und Bepflanzung kosten 70 000 Franken. (St. Galler Tagblatt, 12.04.2010: 36; St. Michael erhält einen Rebberg)

sind (15) und (16) häufiger, (14) und (17) seltener. Unsere Hypothese und damit auch die These von Mackowiak (2008: 47) hat sich bestätigt: In Sätzen ohne Artikelellipse im koordinierten Subjekt ist nach der Grundregel (2.1, *number resolution*) der Plural (15) signifikant überrepräsentiert und der Singular (14) signifikant unterrepräsentiert. In Sätzen mit Artikelellipse (16, 17) kehren sich die Verhältnisse um. Dieser Ellipseneffekt ist wie der Einfluss des Abstraktionsgrades der Nomen im Subjekt höchstsignifikant. Außerdem ist davon auszugehen, dass der Ellipseneffekt im Vergleich zum Effekt des Abstraktionsgrades stärker ist.

Dieser formale Zusammenhang zwischen dem Artikelgebrauch im Subjekt und dem Numerus des finiten Verbs lässt sich semantisch interpretieren: Die Artikelellipse im Subjekt (16, 17) und der Singular des finiten Verbs (14, 16) deuten beide darauf hin, dass eine Aussage über eine einzige Entität gemacht werden soll – auch dann, wenn die Semantik der Nomen dies nicht erzwingt, wenn also der Ersatz des koordinierten Subjekts durch ein Pronomen im Plural möglich wäre. Entsprechend: Selbst wenn der Ersatz des koordinierten Subjekts durch ein singularisches Pronomen möglich ist, kann sowohl der wiederholte Artikel (14, 15) als auch das finite Verb im Plural (15, 17) anzeigen, dass es sich um eine Aussage über zwei Entitäten handeln soll. Kombinationen, bei denen Artikelgebrauch und Verbnumerus in dieselbe Richtung wirken (15, 16), werden bevorzugt, bei unbelebten Abstrakta fast so konsequent wie bei *committee nouns* und Konkreta.

Literaturverzeichnis

Behaghel, Otto (1928): Deutsche Syntax III. Heidelberg: Winter (= Germanische Bibliothek: Abteilung 1, Sammlung germanischer Elementar- und Handbücher: Reihe 1, Grammatiken, 10,3).

Bubenhofer, Noah/Konopka, Marek/Schneider, Roman (2014): Präliminarien einer Korpusgrammatik. Unter Mitarbeit von Caren Brinckmann, Katrin Hein und Bruno Strecker. Tübingen: Narr (= Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache 4).

Cohen, Ayala (1980): On the graphical display of the significant components in two-way contingency tables. In: Communications in Statistics – Theory and Methods 9 (10), S. 1025–1041. DOI: https://doi.org/10.1080/03610928008827940.

Connexor Oy (2011–2016): Machinese Phrase Tagger. https://www.connexor.com/nlplib/?q=mpt (25.01.2017).

Corbett, Greville G. (2000): Number. Cambridge: Cambridge University Press (= Cambridge textbooks in linguistics).

Dammel, Antje (2015): One plus one make(s) – what? Determinants of verb agreement in German NP+NP coordination – A diachronic approach. In: Fleischer, Jürg/Rieken, Elisabeth/Widmer, Paul (Hg.): Agreement from a diachronic perspective. Berlin: de Gruyter Mouton (= Trends in linguistics Studies and monographs 287), S. 287–326.

Dobson, Annette Jane (1990): An Introduction to Generalized Linear Models. London: Chapman and Hall.

Donalies, Elke (2011): Korrespondenz zwischen koordinierten Subjekten und finitem Verb (grammis). https://grammis.ids-mannheim.de/systematische-grammatik/1625, zuletzt aktualisiert am 01.03.2011, zuletzt geprüft am 23.02.2018.

[Duden – Die Grammatik 2016] = Wöllstein, Angelika (Hg.): Duden – Die Grammatik (2016). Bibliographisches Institut. 9., vollständig überarbeitete und aktualisierte Auflage. Berlin: Dudenverlag (= Duden – Deutsche Sprache in 12 Bänden 4).

[Duden – Zweifelsfälle 2012] = Dudenredaktion (2012): Dudenband 9 – Richtiges und gutes Deutsch. Das Wörterbuch der sprachlichen Zweifelsfälle. 7., vollständig überarbeitete Auflage. Unter Mitarbeit von Peter Eisenberg und Jan Georg Schneider. Mannheim: Bibliographisches Institut (= Duden – Deutsche Sprache in 12 Bänden 9).

[Duden – Zweifelsfälle 2016] = Hennig, Mathilde (Hg.) (2016): Das Wörterbuch der sprachlichen Zweifelsfälle. Richtiges und gutes Deutsch. 8., vollständig überarbeitete und erweiterte Auflage. Berlin: Bibliographisches Institut (= Duden – Deutsche Sprache in 12 Bänden 9).

Eisenberg, Peter (2013): Grundriss der deutschen Grammatik. Unter Mitarbeit von Rolf Thieroff. 4., aktualisierte und überarbeitete Auflage. 2 Bände. Stuttgart, Weimar: Verlag J. B. Metzler.

Engel, Ulrich (1996): Deutsche Grammatik. 3., korr. Aufl. Heidelberg: Groos.

Findreng, Ådne (1976): Zur Kongruenz in Person und Numerus zwischen Subjekt und finitem Verb im modernen Deutsch. Oslo: Universitetsforlaget (= Germanistische Schriftenreihe der norwegischen Universitäten und Hochschulen 5).

Friendly, Michael (1992): Graphical Methods for Categorical Data. Paper presented at the SAS SUGI 17 Conference, April, 1992. http://www.math.yorku.ca/SCS/sugi/sugi17-paper.html (25.01.2017).

Gunkel, Lutz, et al. (Hg.) (2016): Grammatik des Deutschen im europäischen Vergleich. Das Nominal. Berlin: de Gruyter Mouton (= Schriften des Instituts für Deutsche Sprache 14).

Hansen-Morath, Sandra/Schneider, Roman/Schmitz, Hans-Christian/Wolfer, Sascha (im Druck): KoGra-R: Standardisierte statistische Auswertung von Korpusrecherchen. In: Fuß, Eric/Konopka, Marek/Wöllstein, Angelika (Hg.): Grammatik im Korpus [Arbeitstitel].

Hastie, Trevor/Pregibon, Daryl (1992): Generalized linear models. In: Chambers, John M./Hastie, Trevor (Hg.): Statistical Models in S. Pacific Grove, Kalifornien: Wadsworth & Brooks/Cole, S. 195–247.

Helbig, Gerhard/Buscha, Joachim (1993): Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. 15., durchges. Aufl. Leipzig: Langenscheidt Verl. Enzyklopädie.

Hennig, Mathilde (2010): Aggregative Koordinationsellipsen im Neuhochdeutschen. In: Arne Ziegler und Christian Braun (Hg.): Historische Textgrammatik und historische Syntax des Deutschen. Traditionen, Innovationen, Perspektiven. Berlin, New York: De Gruyter, S. 937–963.

Hennig, Mathilde (2015): Explizite und elliptische Junktion in der Attribution: Eine Bestandsaufnahme. In: Hennig, Mathilde/Niemann, Robert (Hg.): Junktion in der Attribution. Ein Komplexitätsphänomen aus grammatischer, psycholinguistischer und praxistheoretischer Perspektive (= Linguistik – Impulse & Tendenzen 62), S. 21–84.

Hennig, Mathilde (2017): Grammatik und Variation im Spannungsfeld von Sprachwissenschaft und öffentlicher Sprachreflexion. In: Konopka, Marek/Wöllstein, Angelika (Hg.): Grammatische Variation. Empirische Zugänge und theoretische Modellierung (= Jahrbuch des Instituts für Deutsche Sprache), S. 23–45.

Heycock, Caroline/Zamparelli, Roberto (2005): Friends and Colleagues. Plurality, Coordination, and the Structure of DP. In: Natural Language Semantics 13 (3), S. 201–270. DOI: 10.1007/s11050-004-2442-z.

Hoffmann, Ludger (2016): Deutsche Grammatik. Grundlagen für Lehrerausbildung, Schule, Deutsch als Zweitsprache und Deutsch als Fremdsprache. 3., neu bearb. und erw. Auflage. Berlin: Schmidt.

Institut für Deutsche Sprache (2014): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2014-I (Release vom 15.04.2014). http://www1.ids-mannheim.de/direktion/kl/projekte/korpora/releases.html (07.12.2016).

Institut für Deutsche Sprache (2015): KoGra-R: Standardisierte statistische Verfahren für korpusbasierte Häufigkeiten. http://kograno.ids-mannheim.de/index.html (25.01.2017).

Institut für Deutsche Sprache (2016a): Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2016-I (Release vom 31.03.2016). http://www1.ids-mannheim.de/direktion/kl/projekte/korpora/releases.html (07.12.2016).

Institut für Deutsche Sprache (2016b): Korpora. http://www.ids-mannheim.de/cosmas2/projekt/referenz/korpora.html, zuletzt aktualisiert am 23.11.2016 (zuletzt geprüft am 01.12.2016).

Mackowiak, Klaus (2008): Die 101 häufigsten Fehler im Deutschen und wie man sie vermeidet. Orig.-Ausg., 3., aktualisierte, neu bearb. und erw. Aufl. München: Beck (= Beck'sche Reihe 1667).

McCullagh, Peter / Nelder, John A. (1989): Generalized linear models. 2. Aufl. London: Chapman and Hall (= Monographs on statistics and applied probability 37).

Meyer, David / Zeileis, Achim / Hornik, Kurt (2005): The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd. http://epub.wu.ac.at/480/, zuletzt aktualisiert am 09.10.2013 (25.01.2017).

Schmid, Helmut (1995): Improvements in Part-of-Speech Tagging with an Application to German. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger2.pdf.

Schrodt, Richard (2005): Kongruenzprobleme bei Subjekt und Prädikat. Die Termqualität geht vor. In: Eichinger, Ludwig M./Kallmeyer, Werner (Hg.): Standardvariation. Wie viel Variation verträgt die deutsche Sprache? Berlin, New York: Walter de Gruyter (= Jahrbuch des Instituts für Deutsche Sprache 2004), S. 231–246.

Zifonun, Gisela / Hoffmann, Ludger / Strecker, Bruno (1997): Grammatik der deutschen Sprache. 3 Bände. Berlin, New York: De Gruyter (= Schriften des Instituts für Deutsche Sprache 7).

Tom Bossuyt, Ludovic De Cuypere, Torsten Leuschner

Emergence Phenomena in German *W-immer/auch*-Subordinators

Abstract The present study is concerned with the distributional patterns of the irrelevance particles *immer* 'ever' and *auch* 'also' in German universal concessive conditionals and free relatives (e.g. *was immer er auch sagt* 'whatever he says'). Whereas irrelevance is conveyed by a single element in a fixed position in languages like English *(-ever)*, *immer* and *auch* occur in multiple positions and combinations. Following the example of Leuschner (2000), the distribution of particles and their combinations is documented and explained using functional motivations. Compared with Leuschner (2000), however, the present study is based on a much larger sample of 23,299 clauses with the *W*-words *was* and *wer* (incl. their inflected forms) from the *DeReKo*-corpus, allowing for a far more detailed statistical analysis. Special attention is devoted to the distribution of *immer* and *auch* (including their combinations) in full subordinate clauses vs. elliptically reduced forms, and to the nature of the resulting patterns as a case of emergent grammar.

Keywords Concessive conditionals; irrelevance; particles; subordinators; emergent grammar; corpus study

1 Introduction

Following König (1986), it has become customary to analyse adverbial subclauses like those in (1a.–c.) as different subtypes of *concessive conditionals*:

(1) a. Universal concessive conditional
However much financial support we get, we will go ahead with our project.
b. Alternative concessive conditional
Whether we get financial support or not, we will go ahead with our project.

c. Scalar concessive conditional
Even if we do not get financial support, we will go ahead with our project.
(cf. Haspelmath/König 1998:563)

The term "concessive conditional" (henceforth: CC) has been adopted by other researchers (e.g. Breindl 2014) and even found its way into some reference works (e.g. Zifonun et al. 1997). Despite their heterogeneous form in some languages (including English and its relatives, Haspelmath/König 1998), all CCs express the same basic conditional meaning (cf. König 1986, Leuschner 2006, d'Avis 2016):

(2) a. if $\{p_1$ or p_2 or p_3 or ...$\}$, then q
b. if p_n, then normally not q

Instead of just one antecedent value (if p then q), the various subtypes use different strategies to invoke a multiplicity of antecedent values (if p_x then q), whose individual truth values are irrelevant to the truth value of the consequent q in the apodosis. The values form a set which is partially ordered along some relevant parameter (i.e. a partially ordered set or 'poset', cf. Neggers/Kim 1998), hence the protasis typically contains a contextually extreme antecedent condition p_n, under which q would not normally be expected to be true, as suggested by (2b.) (König 1986:234). For example, the subclauses in (1a.-c.) all invoke a set of values along the parameter 'amount of funding obtained'. (1a.) does so by means of a *WH-ever*-type quantificational expression, (1b.) by means of a disjunction naming the two endpoints of the scale, and (1c.) by marking one of the endpoints (failure to obtain funding) as a particularly informative value by means of the scalar focus particle *even*. As projects are normally cancelled in the absence of funding, all three subtypes assert with particular force the continuation of the project regardless of financial circumstances.

This paper is concerned with the first of the three subtypes (henceforth: UCCs) in German. While introducing the label *universal concessive conditional*, König/Eisenberg (1984) admitted that the relevant quantificational strategy is in fact quite different from standard universal quantification (König/Eisenberg 1984: 315). Instead, UCCs "signal a free choice in the selection of values for a variable in the protasis" (König 1986: 231) and are therefore more reminiscent of *any* than of *every* or *all*.

In English UCCs, free-choice quantification is invariably marked by a standard item, viz. *-ever*, in a fixed position, viz. attached to the *WH*-word.[1] German,

1 An exception is *WH-so-ever* (e.g. *whatsoever*), which will be briefly discussed below.

by contrast, has two corresponding items, viz. *immer* 'ever'[2] and *auch* 'also'. As far as their use as free-choice markers in UCCs and related constructions (cf. below) is concerned, *-ever*, *immer* and *auch* will henceforth be referred to as *irrelevance particles* (cf. Leuschner 2000: 344). Whereas *-ever* as irrelevance particle fails to show any positional variability across the clause, *immer* and *auch* may occur in different positions, either alone or combined, as shown in (3):[3]

(3) a. *Was immer* er sagt, keiner hört ihm zu.
b. *Was* er *auch* sagt, keiner hört ihm zu.
c. *Was immer* er *auch* sagt, keiner hört ihm zu.
d. *Was immer auch* er sagt, keiner hört ihm zu.
e. *Was auch immer* er sagt, keiner hört ihm zu.
f. *Was* er *auch immer* sagt, keiner hört ihm zu.
'Whatever he says, nobody listens to him.'

Furthermore, *immer*, with or without *auch*, is attested marginally after pronominal subjects (cf. below for figures), while *auch* is sometimes placed in front of lexical subjects:

(4) a. *Was* er *immer* sagt, keiner hört ihm zu.
b. *Was* er *immer auch* sagt, keiner hört ihm zu.
c. *Was auch* der alte Mann sagt, keiner hört ihm zu.
'Whatever he/the old man says, nobody listens to him.'

For decades, descriptive grammars of German have tended to overlook and/or simplify these positional and combinatorial patterns. Most suggest vaguely that either *immer* or *auch* is obligatory, while the other can be omitted (cf. Bossuyt 2016: 49f. for a more detailed survey). In response to this situation, which remains essentially unchanged today, Leuschner (2000) first investigated the patterns and frequencies of *immer/auch* in UCCs in 104 examples gleaned from the *Mannheimer Korpus* (ca. 2.2 million tokens in total). His main conclusions were

2 German *immer* is a partial cognate of English *ever* through the initial *i-* (Middle High German *ie* in *ie-mêr*, cf. Modern High German *je* 'ever'), which is cognate with the initial *e-* of *ever* (Old English *æ-fre*, Leuschner 1996). *Immer* had free-choice 'ever' as one of its standard temporal readings in earlier German (ibd.) and continues to retain a non-universal, *ever*-like reading in combination with adjectives even today, as e.g. in *immer größer* 'ever greater'. The free-choice meaning of *immer* goes back historically to the temporal 'ever'-reading, but like *-ever*, *immer* has lost all temporal force in UCCs (Leuschner 1996: 481).

3 All *W*-words (e.g. *was*) and irrelevance particles in example sentences are italicised.

(i) that *immer* and *auch* show complementary positional tendencies: *immer* is invariably adjacent to the clause-initial *W*-word, while *auch* tends strongly (though not necessarily) to occur towards the clause-final verb phrase;
(ii) that *immer* and *auch*, when used in the same clause and in this order, retain their individual positional preferences, hence the subject – and possibly some other constituent – may be placed in between *immer* and *auch*, as in (3c.), creating an *immer* (…) *auch* pattern;
(iii) that the combination of *auch* with *immer*, in this order, does not allow other elements in the clause to intervene and that this pattern, represented simply as *auch immer*, also shows a "preference for shorter and elliptically reduced subclauses" (Leuschner 2000: 353).

Compared with the *Mannheimer Korpus* (which dates from the 1960s), corpus sizes have increased vastly in recent years, creating unprecedented opportunities for analysis. A prominent example is the *Mannheimer Korpus* itself, which has since been included in the much larger *Deutsches Referenzkorpus* (*DeReKo*; Kupietz et al. 2010, Kupietz/Lüngen 2014). On the quantitative side, our paper draws on the *DeReKo* in a partial replication of Leuschner's (2000) study, using a much-expanded sample (with some inevitable restrictions of its own) and a more sophisticated statistical methodology. On the qualitative side, we develop for the first time the hypothesis that the positional and distributional patterns of *immer* and *auch* represent a snapshot of the long-term emergence of irrelevance marking as a subsystem of modern German. German combinations of clause-initial *W*-words with *immer* and/or *auch*, we argue, form a long-term building-site of grammaticalisation ("Grammatikalisierungsbaustelle", Leuschner 2006, cf. Nübling 2005) whose completion will remain uncertain until *immer* is finally reanalysed as part of the *W*-phrase and univerbated with the *W*-word. While this happened to the English *-ever* several centuries ago (Leuschner 2006:135f.), such a step continues to look unlikely in German for the foreseeable future.

2 Methodology

DeReKo, the corpus used for the present study, is the main reference corpus for modern German, containing ca. 42 billion words of running text as of February 3, 2018.[4] Based on a broad sample of written genres, including fiction, most texts are from printed news media; Wikipedia articles and discussions have recently been included, as have parliamentary minutes (Kupierz/Lüngen 2014, cf. Scherer 2014:83). As in Leuschner (2000), the search was targeted at *W*-words followed by

4 http://www1.ids-mannheim.de/kl/projekte/korpora/, last accessed February 25, 2018.

immer and/or *auch*, but unlike Leuschner (2000), who searched for all *W*-words, including *wann* 'when', *wo* 'where' etc., we restricted our query, for practical reasons, to *was* 'what' and the paradigm of *wer* 'who' (i.e. nominative *wer*, genitive *wessen*, dative *wem* and accusative *wen*; cf. Thieroff 2011). Before the search, decisions had to be taken on the distance operators in the search queries, i.e., the distance in number of words between *was* or *wer* (incl. inflectional forms) and *immer/auch*. Taking into account Leuschner's (2000) conclusions on the positional tendencies of *immer* and *auch*, only instances of *immer* immediately following the *W*-word were included (i.e. the distance operator was set to 1), whereas a distance operator of 4 words was applied with *auch*.[5] A total of 48,464 tokens were then exported from *DeReKo* on December 23rd, 2015. A preliminary analysis of *was*, which alone yielded 8,734 tokens, can be found in Bossuyt (2016) and has been incorporated into the results below. 5,268 additional tokens were exported on November 11th, 2016, with *immer* immediately preceded by a 3rd person singular pronoun which was in turn preceded immediately by the *W*-word (e.g. *was es immer*),[6] bringing the total of exported tokens to 53,732. All tokens were analysed manually to check whether *immer* and *auch* did indeed function as irrelevance particles – after all, *immer* can be a temporal adverb and *auch* can be an additive focus particle – and to remove doubles containing the particles in combination. This brought the final sample to 23,299 tokens.

Not all these tokens represent prototypical UCCs like those mentioned above. Some are non-specific free relatives (henceforth: NFRs) as in (5):

(5) *Wer immer* sich angesprochen fühlt, ist dazu eingeladen. (A99/FEB.12351)
'Whoever feels addressed, is invited.'

Whereas the protasis in UCCs functions as a loose adjunct of the apodosis, NFRs typically function as embedded arguments in the matrix clause (Leuschner 2005), e.g. as its subject in (5). However, the distinction between UCCs and NFRs is not clear-cut (cf. Leuschner 2005:59–62), and since both types constitute genuine subclauses with a clause-initial *W*-word followed by one or more irrelevance particles, we will jointly designate all *W immer/auch*-constructions which

5 A distance operator of 3 was selected for *wessen* (…) *immer* because this *W*-word can modify NPs. For *W* (…) *auch*, a distance operator of 4 seemed to be the most practical solution: clauses with subjects consisting of a determiner, adjective and noun could still be found, without the distance between the *W*-word and *auch* being too large, causing an undesirably large number of invalid instances to be found, e.g. where *auch* occurs in the apodosis or in the next sentence.

6 We are grateful to Dr. Eric Fuß (IDS Mannheim) for suggesting this strategy. Although the search yielded only a small number of new tokens with *immer* as irrelevance particle, our database did become more comprehensive as a result.

function as subclauses as *primary constructions* (as opposed to secondary constructions, cf. below).

Primary constructions are analysed using Leuschner's (2000) version of the Topological Field Model for German clause structure (cf. Wöllstein 2014) as can be seen in Table 1a.

Table 1a: Leuschner's (2000:345) version of the Topological Field Model, exemplified by (4c).

pre-field	left bracket	middle field			right bracket	post-field
W	–	II	S	IV	V	–
was	–	*immer*	er	*auch*	sagt	–

While the *W*-word occupies the pre-field, leaving the left bracket unoccupied in Standard German (Wöllstein 2014: 32–37), the middle field is divided into a field for the subject of the subclause (S) and two fields which may be occupied by irrelevance particles: field II to the left of S and field IV to the right of S (Leuschner 2000:345). As usual in German subclauses, the VP occupies the right bracket (V), followed by the post-field, which is empty as a default.

The topological model in Table 1a only makes sense if the *W*-word is not the subject of the subclause. When the *W*-word is the subject, there is no need to split up the middle field, and the latter is then simply called II/IV (Leuschner 2000: 346) as can be seen in Table 1b:

Table 1b: Leuschner's (2000: 346) version of the Topological Field Model, exemplified by (5).

pre-field	left bracket	middle field	right bracket	post-field
W	–	II/IV	V	–
wer	–	*immer* sich	angesprochen fühlt	–

While these two models fit nearly four fifths of all tokens, 4,926 (21.14 %) do not fit either model. The reason is that they are derived historically from primary constructions by ellipsis and reduced to a *W*-word + irrelevance particle(s) combination (cf. Breindl 2014: 98of., Leuschner 2013: 57, Waßner 2006: 386f.). We label them *secondary constructions*. They may function as:

(6) indefinite pronouns (cf. Haspelmath 1997: 139, 16of.):
Ein Appell an *wen auch immer*, der sich verantwortlich fühlt.
(U08/JUL.03097)
'A call to anyone (lit. whoever) who feels responsible.'

(7) discourse markers (more usually *wie auch immer* 'however', Leuschner 2000: 352):
Doch *was auch immer*: Ein Crash ist trotzdem jederzeit möglich. (SOZ06/OKT.04291)
'But whatever: a crash is nevertheless a possibility at all times.'

(8) "general extenders" (Overstreet 1999):
Ich bete mit Ihnen zu Gott – oder zur Göttin oder *wem auch immer* (PBE/W15.00007)
'I pray with you to God – or to the Goddess or whoever.'

Since irrelevance particles show a strikingly different distributional behaviour in primary and secondary constructions, we distinguish between primary and secondary constructions in the sections which follow. Section 3 presents our results, first regarding the former (3.1), then the latter (3.2). After sketching the diachronic emergence of the particles' positional tendencies (section 4.1), we then similarly analyse our results first with respect to primary constructions (sections 4.2–5.1), then to secondary constructions (section 5.2), before turning to the conclusion (section 6).

3 Basic distributional patterns

3.1 Primary constructions

Table 2a presents the distribution of irrelevance particles in primary constructions in which the *W*-word is not the subject of the subclause.[7] An example from the corpus for each type is given in (9).

(9) a. *Was auch* die Gründe sein mögen, nur jammern […] hilft auch nicht weiter. (A01/OKT.32079)
'Whatever the reasons may be, just complaining won't help either.'
b. *Wen auch immer* man fragt: Esel finden alle irgendwie klasse. (U06/JUN.00549)
'Whoever you ask: everyone thinks donkeys are great somehow.'
c. *Wer immer auch* die Täter sind, […], sie müssen sich vorsehen. (SOZ10/APR.03622)
'Whoever the perpetrators are, they have to watch out.'

7 Note that the left bracket and the post-field are omitted from this and the following tables, as they are irrelevant to the particles' distribution.

Table 2a: Distribution of irrelevance particles in subclauses where W ≠ S.

	W	II	S	IV	V	#	%
(a)	W	*auch*	S	–	V	22	0.24 %
(b)	W	*auch immer*	S	–	V	954	10.53 %
(c)	W	*immer auch*	S	–	V	149	1.64 %
(d)	W	*immer*	S	–	V	6,075	67.05 %
(e)	W	*immer*	S	*auch*	V	1,005	11.09 %
(f)	W	–	S	*auch*	V	647	7.14 %
(g)	W	–	S	*auch immer*	V	154	1.70 %
(h)	W	–	S	*immer auch*	V	15	0.17 %
(i)	W	–	S	*immer*	V	39	0.43 %
						9,060	100.00 %

d. *Was immer* sie tun, Maitressen haben einen schlechten Ruf. (U14/APR.01817)
'Whatever they do, mistresses have a bad reputation.'
e. Doch *was immer* er *auch* tut, es reicht nicht. (T13/NOV.02370)
'But whatever he does, it is not enough.'
f. Mit *wem* ich *auch* rede, überall höre ich dasselbe. (PBE/W14.00030)
'Whoever I talk to, I hear the same everywhere.'
g. *Wessen* Socke das *auch immer* ist, es wird langsam langweilig. (WDD11/P57.49531)
'Whoever's sock that is, it's beginning to get boring.'
h. Zeitgemäße Dienstvereinbarungen, *was* das *immer auch* heißen möge. (PNO/W15.00042)
'Contemporary service contracts, whatever that may be.'
i. *Wer* es *immer* wissen könnte, M. M. weiß es nicht. (T05/APR.02136)
'Whoever might know about it, M. M. does not know about it.'

Overall, the preferred position of irrelevance particles is clearly in field II rather than in field IV. 79.47 % of all tokens have (all) their irrelevance particles in this field II (= types a.–d.), only 9.44 % have it/them in field IV (= types f.–i.). The latter is less than the 11 % of tokens which have particles in both II and IV (= type e.); if we add this type to those of the first, the cumulative proportion of particles in field II amounts to 90.56 % of all tokens. The language-specific distribution of particles in German thus mirrors the overall tendency for irrelevance

particles in Standard Average European to immediately follow, or be suffixed to, the w-word[8] (Haspelmath/König 1998: 609). The other option, viz. "clause-internal" placement further to the right, is a minority option cross-linguistically (ibd.), and so it is in German.

Empirically speaking, this distributional pattern is due to the very high proportion of tokens with *immer* (67.05 %), and to a lesser extent *auch immer* (10.53 %), in field II. Other relatively frequent variants are *immer* (…) *auch*, which has *immer* in field II and simultaneously *auch* in field IV (11.09 %), and *auch* alone in field IV (7.14 %). All other variants account for less than 2 % each, or about 4.18 % in total. This distribution deviates somewhat from Leuschner's (2000) findings, the most striking differences being the much higher proportion of *immer* in our data (6,075 out of 9,060 tokens or 67.05 % compared to just 34 out of 92 tokens or 36.96 % in Leuschner 2000: 348) and the much lower proportion of *auch* (647 out of 9,060 tokens or 7.14 % compared to 38 out of 92 tokens or 41.3 % in ibd.). A two-tailed two-proportions Z-test suggests that these deviations are significant ($p < 0.0001$), possibly reflecting differences between the corpora. The results do, however, square with the particle-specific positional tendencies observed by Leuschner (2000): *immer* shows a very strong tendency to occupy field II (6,075 out of 6,114 tokens = 99.36 %), and *auch* has a clear preference for field IV (647/669 = 96.71 %).

Particle combinations have positional tendencies of their own. *Immer* (…) *auch* mostly straddles the subject field (1,005/1,169 = 85.97 %), so that each of its constituent particles occupies its own field of preference (*immer* II, *auch* IV). By contrast, *auch immer* is never broken up by any constituent and shows a strong left-leaning tendency (954/1,108 = 86.1 %). Using the terminology suggested by Thurmair (1989: 290) for combinations of modal particles, *auch immer* therefore qualifies as a "closed" particle combination, i.e. one that behaves like a single, complex particle, and *immer* (…) *auch* as an "open" combination of two individual particles that may or may not be mutually adjacent.

Finally, Table 2b shows the distribution of irrelevance particles in primary constructions in which the *W*-word is the subject of the subclause. An example from the corpus for each type is given in (10).

8 The term w-word, with w rendered as a non-italicised small capital, is used here as a language-independent designation. Regular, italicised capitals are used for language-specific categories, i.e. *W*-words in German and *WH*-words in English. According to this convention, English *how* is subsumed under *WH*-words despite its spelling; however, *how* does not in fact play a role in the present study.

Table 2b: Distribution of irrelevance particles in subclauses where W = S.

	W	II/IV	V	#	%
(a)	W	*auch*	V	79	0.85 %
(b)	W	*auch immer*	V	1,295	13.91 %
(c)	W	*immer auch*	V	640	6.87 %
(d)	W	*immer*	V	7,299	78.37 %
				9,313	100.00 %

(10) a. Denn *was auch* passiert: Freilichtspiele sind immer ein Erlebnis. (M01/JUN.44510)
'For whatever happens: open-air theatre is always a great experience.'
b. *Was auch immer* passiert, es muss schnell geschehen. (LTB11/JUN.00726)
'Whatever happens, it has to happen fast.'
c. *Was immer auch* passiert, Gott will, daß wir glücklich sind. (O95/JAN.07794)
'Whatever happens, God wants us to be happy.'
d. *Was immer* passiert, wir sind bereit zu kämpfen. (A99/FEB.11037)
'Whatever happens, we are prepared to fight.'

Compared with Table 2a, the proportions of *immer* and *auch immer* are significantly higher, while the proportions of *auch* and *immer auch* are significantly lower (both based on a two-tailed two-proportions Z-test, $p < 0.001$).

3.2 Secondary constructions

In secondary constructions, irrelevance particles are distributed very differently compared to primary constructions, as shown by Table 3.

Table 3: Distribution of irrelevance particles in secondary constructions.

	immer	immer auch	auch immer	auch	total
#	399	18	4,485	24	4,926
%	8.10 %	0.37 %	91.05 %	0.49 %	100 %

Whereas *immer* is the most frequent particle in primary constructions, it plays a strikingly minor role in secondary constructions (8.1 %). Instead, *auch immer* is clearly dominant in secondary constructions (91.05 %). *Auch immer* is also the only particle (or particle combination) that prefers secondary constructions, as 4,485 out of 6,888 tokens with *auch immer* (= 65.1 %) occur in secondary constructions. For all other irrelevance particles or particle combinations, by contrast, use in secondary constructions is dispreferred (*immer*: 399/13,812 = 2.89 %; *immer*

auch: 18/1,827 = 0.99 %; *auch*: 24/772 = 3.11 %). This confirms the "preference for shorter and elliptically reduced subclauses", i.e. secondary constructions, found by Leuschner (2000: 353) with *auch immer*.

4 Irrelevance marking as an emergent system

4.1 Historical background

As suggested earlier (cf. chapter 1. *Introduction*), the positional and distributional patterns of *auch* and *immer* and their combinations can be read as a snapshot of the long-term emergence of irrelevance marking in modern German. This process follows historically from the simplification of the *so W so* irrelevance marking construction that Old High German inherited from ancient West Germanic (Leuschner 2006: 134; Lühr 1998). Here are examples of *so W so* in Old High German and of its Old English counterpart, *swa WH swa*:

(11) a. *So wér so* ist fona wáre, ther hórit mir io sáre.
'Whoever is from the truth, he always obeys me immediately.'
(cited in Leuschner 2001:16)

b. *Swa hwylc swa* næfð, þæt he wene þæt he hæbbe, him bið afyrred.
'Whoever has nothing, what he thinks he has will be taken away from him.'
(cited in ibd.:15)

Given the semantic opacity of SO ... SO as an irrelevance marking strategy[9] and the fact that both SO were unstressed (Lühr 1998), it is no wonder that the simplification of SO W SO and the replacement of SO ... SO with semantically more transparent strategies began with the omission of one SO (see Leuschner 2001 for a survey of this process in a Germanic-wide context, and 2006: 134–140 for a summary in English). In Old English, it was the left-hand *swa* that was dropped first, and the adverb *æfre* 'ever' already began to be added to support the quantificational effect:

9 By analogy with the convention established in footnote 8, we use non-italicised, small-capitals SO as a language-independent designation which subsumes language-specific *swa* and *so*. By analogy with modern English *how*, *WH*-words include the Old English predecessors of modern *who*, *what* etc. such as *hwa* 'who' despite their spelling.

(12) Luue ðine nexte al swa ðe seluen, *hwat* manne *swa* he *æure* bie!
'Love thy neighbor like thyself, whatever man he be!'
(cited in Leuschner 2006:135)

After it was introduced in what we identified above as field IV, i.e. in the typical position of adverbs, *æfre* was reanalyzed as a quantificational particle and gradually moved left towards the *WH*-word. While the surviving right-hand *swa* (> *so*) could still serve as sole irrelevance particle for several centuries, *æfre* (> ME. *æure* > *ever*) became more and more obligatory; with both *so* and *ever* increasingly cliticised to the *WH*-word, *so* was eventually squeezed out, surviving today almost only in the postnominal Negative Polarity Item *whatsoever* as in *no idea whatsoever* 'no idea at all' (Leuschner 2001:9). In all other cases, *so*-less *WH-ever* is now the only remaining option.

In contrast to English, the corresponding changes in German began with the initial loss of right-hand *so*; left-hand *so* was then weakened to *se* in Middle High German, later cliticised as *s-* to the *W*-word (as e.g. in *swer* 'whoever') and eventually lost altogether, causing the erstwhile *s-W*-words to collapse with the bare *W*-words during the fourteenth century (Leuschner 2006: 135). By this time, *iemer* 'ever' (> *immer*) and *ouch* 'also' (> *auch*) had been introduced as alternative irrelevance markers, along with several other particles which later disappeared again:

(13) a. er sol swern, dise stat ze behaltene, *swâ* er *iemer* allermeist kan
'he shall swear to keep this place wherever he can'
(cited in Leuschner 2000: 349)
b. diu schamt sich des, *swâ iemer* wîbes scham geschiht
'she is ashamed of it, wherever dishonour happens to a woman'
(cited in Leuschner 2006: 135)
c. *swaz ouch* mir dâ von geschiht
'whatever happens to me as a consequence'
(cited in Leuschner 2006: 136)

In contrast to English, where some irrelevance marking was always in place, use of *immer* and/or *auch* was still optional by the early 19th century, as shown by this example by J. W. Goethe (1749–1832):

(14) *Was* ich thue, *was* ich lasse; / Nur ein unbestimmt Verlangen / Fühl' ich, das die Brust durchglüht.
'Whatever I do, whatever I do not do; all I feel is an uncertain desire glowing in my breast.'
(cited in Leuschner 2006:136)

Not until the twentieth century did the presence of at least one irrelevance particle become mandatory, as it is today (d'Avis 2016: 277). Nor did the positional tendencies of *immer* and *auch* become clear until well into the nineteenth century (Leuschner 2006: 136), as again suggested by examples from the works of Goethe, who was still able to position *immer* in field IV:

(15) Und man kommt in's Gered', *wie* man sich *immer* stellt.
'And one becomes the subject of gossip, however one (lit.: how one ever) positions oneself.'
(cited in Goethe's *Faust I*, line 3201)

Immer has since replicated the leftward shift of *ever* (cf. above), albeit less consistently (cf. above); it took several centuries longer than *ever* to do so and has so far failed to reach the corresponding conclusion (Leuschner 2006: 136). And of course, the picture is complicated further by the presence of *auch*, which has been undergoing its own (partial) shift in the reverse direction from field II to field IV, and often combines with *immer* in fields II and IV.

4.2 Disambiguation

The emergence of a separate paradigm of *WH-ever* conjunctions in English bears many hallmarks of grammaticalisation (cf. Lehmann 1995) such as semantic bleaching of *ever*, increased condensation through *WH*-adjacency and cliticisation, as well as obligatorification. With this highly advanced process as background, the question arises what, on the one hand, has been driving the corresponding process in German and what, on the other hand, been hindering its completion.[10]

10 An anonymous reviewer suggests that the very presence of irrelevance marking in German is redundant given the characteristic disintegration of the clause complex, as in (3) and (4) above. In this view, the loose adjunction of the (sentence-initial) protasis to an apodosis with separate V2 word order (cf. König/van der Auwera 1988) is characteristic enough to serve as a kind of irrelevance marking in its own right. This would hardly be an effective strategy, however, as the listener would have to wait until the onset of the apodosis in order to identify retrospectively the intended interpretation of the protasis. Syntactic disintegration does not come into play at all when the protasis is non-sentence-initial, and in cases where the protasis functions as an NFR as in example (5) above, or in some intermediate function (Leuschner 2005), it does not offer sufficient clues, either. We therefore continue to believe that irrelevance marking at the level of the subclause is functionally well-motivated in its own right and in no way redundant, as indeed suggested by the systemic dynamism that is the object of our investigation.

According to Leuschner (2000: 347), the above-mentioned positional change of *immer* (and earlier of *ever* in English) towards *W*-adjacency has been motivated by disambiguation. Whereas *immer* unambiguously functions as an irrelevance particle adjacent to the *W*-word, as in (16), it is prone to be mistaken for the temporal adverb *immer* 'always' when placed near the verb, as in (16)'.

(16) *Was immer* die drei Musiker spielen […] (A97/MAI.01784)
'whatever the three musicians play'
(16)' *Was* die drei Musiker *immer* spielen […]
'what the three musicians always play'

Just as positioning the irrelevance particle *immer* in field II distinguishes it from the temporal adverb, positioning *auch* in field IV helps keep it distinct from its alternative function as the additive focus particle *auch* 'also, even'. *Auch* is more likely to be read as an irrelevance particle when it is close to the verb as in (17), and more likely to be read as an additive focus particle when it is close to the *W*-word as in (17)'.

(17) *Was* die Mexikaner *auch* anpacken […] (H86/OM3.11688)
'whatever the Mexicans tackle'
(17)' *Was auch* die Mexikaner anpacken […]
'what also/even the Mexicans tackle'

We conclude that the complementary preferences of *immer* and *auch* for fields II and IV, respectively, are brought about by the same functional motivation: disambiguation. With *immer*, disambiguation by adjacency is absolute: if *immer* is adjacent to the *W*-word, it cannot be an adverb and must be read as an irrelevance particle. With *auch*, the disambiguation effect is less inevitable and, in the spoken medium, partly linked to stress: stressed *auch* is more likely to be read as a focus particle than unstressed, regardless of position. In written data like (17) and (17)', whether *auch* is an irrelevance particle or not can only be decided on grounds of context, yet the results are clear, showing irrelevance *auch* being placed overwhelmingly near the verb (cf. Table 2a: 647 times in field IV vs. 22 times in field II).

In view of the clear tendency of *immer* towards *W*-adjacency, it is tempting to conclude that *WH-ever*-like subordinating conjunctions with *immer* (i.e. *wer-immer* 'whoever', *wasimmer* 'whatever' etc.) may be formed at some stage in German in the near future. Unfortunately for this prospect, the required univerbation of *immer* with the *W*-word is unlikely to take place any time soon, given that other material may intervene, either optionally or required between any irrelevance particle and the *W*-word. In cases with optional material intervening,

it is in fact extremely rare to find *immer*. Exceptions like (18) require well-targeted search queries to be identified in *DeReKo*.

(18) *Was* aber *immer* sie zur Rechtfertigung ihrer Versäumnisse vorbringt […] (P93/FEB.05671)
'But whatever she puts forward as a justification of her failures'

By contrast, *auch* or *auch immer* can occur in this position (i.e. field II, but not immediately adjacent to the *W*-word) without problems, as seen in (19)-(21):

(19) *Was* genau *auch* das Problem sein kann […] (NUN12/SEP.01641)
'Whatever exactly could be the problem'

(20) *Wem* aber *auch immer* der schwarze Peter nun zufallen wird […] (RHZ97/APR.01964)
'But whoever will be responsible/to blame'
(lit.: But whoever the black Pete will be passed to)

Even PPs with full lexical NPs are easily allowed between the *W*-word and the particle, as in (21):

(21) Auf *wen* im kommenden Jahr *auch* die Entscheidung fällt […] (RHZ11/JUN.00482)
'Whoever will be chosen in the coming year'
(lit.: Whoever the decision will fall upon in the coming year)

Whereas the intervening material in (19)-(21) is optional, in other cases it is mandatory, as e.g. in combinations of *wie* 'how' + adjective and *welch-* 'which' + NP (Leuschner 2000: 350). While phrases like *however beautiful* and *whichever house* are perfectly grammatical in English, their German equivalents are ungrammatical or at least highly unusual and unattested in our data: *wie* *(*immer*) *schön* ?(*immer*), *welches* *(*immer*) *Haus* ?(*immer*). When *wessen* 'whose' modifies an intervening NP as in (22a.), combinations with *immer* alone are similarly ruled out, while combinations with *auch immer* are allowed. When *wessen* functions as a genitive object, on the other hand, and no material therefore intervenes between it and the particle as in (22b.), *immer* is unproblematic.

(22) a. mit *wessen* Geld *auch immer* [**immer*] sie bezahlt wurden (A10/MAR.05697)
'with whoever's money they got payed'

b. *wessen immer* man mich anklagt (U98/MAR.22976)
'Whatever (some)one accuses me of'

Furthermore, Leuschner (2000:350) suggests that *immer* is unable to combine with complex *W*-words like *woher/wohin* 'where from/to', *womit* 'where-with, i.e. with which/what' etc. which are not part of our present sample. These restrictions have so far kept *immer* from attaining full condensation with the *W*-word to a point where it could be reanalysed as part of the W-field and univerbated with the *W*-word. Even worse for this prospect, the preference of *immer* for strict adjacency to the *W*-word has been hindering, not promoting, its obligatorification (cf. ibd.), as its near-exclusion from other positions in field II has been encouraging the use of *auch* or of combinations of *auch* with *immer* rather than *immer* alone.

4.3 The role of the subject

Another significant factor in the emergence of irrelevance marking in primary constructions is the nature of clause-internal subjects. Leuschner (2000: 350) already notes in passing that *auch* seems to occupy its dispreferred field II only if the subject is a lexical NP, never if it is pronominal. Our data confirm this tendency; indeed they show that it is almost exceptionless. Only one counterexample – with *auch* in field II followed by a pronominal rather than lexical subject – is found in the entire sample:

(23) Der Satz mit C. M. – *wer auch* das sein mag – gefällt mir nicht. (WDD11/B18.96254)
'The sentence with C. M. – whoever that may be – is not something I like.'

Similarly, in those rare cases where *immer* occupies its strongly dispreferred field IV, the subject is invariably a pronoun, never (in our data) a lexical NP.[11] This helps explain why **W auch S immer V* is the only logically possible distributional pattern that is not attested at all in the sample. Not only would *auch* and *immer* both occupy their dispreferred fields, depending on the type of subject, this structure could also produce counterexamples to the tendency for *immer* to occur only with pronominal subjects in field IV and to the tendency (apparently

11 The corresponding Middle High German example cited by Leuschner (2000: 349) in (13a.) has *iemer* in field IV following the pronominal subject *er* 'he'. Further study of historical data is required, but so far we are not aware of any instances from any period where *immer* occupies field IV after a lexical subject.

almost exceptionless) for *auch* to co-occur only with lexical subjects in field II. *Immer auch* never occurs after lexical subjects, just like *immer*.

Before we address potential explanations for these tendencies (cf. below), we emphasize again that the nature of the subject correlates significantly with the possibility for single irrelevance particles to occur in their dispreferred fields. Since **W auch S immer V* is effectively ruled out by a conspiracy of preferences determining the positions of individual particles, these particles will invariably be mutually adjacenct in either field II or IV whenever they occur together in this order, and this must have been a supporting factor in their reanalysis as a single, complex particle (cf. below). By contrast, when *immer* and *auch* occur together in this order, they tend to be pulled apart by their complementary positional preferences, hence there is far less chance for reanalysis to occur.

Let us take a closer look at factors that motivate the choice between the two particle combinations, taking the perspective of field II as suggested by Table 4a.

Table 4a: Types of subject and open/closed combinations of *immer* (…) *auch* in field II. In an open combination, both fields II and IV are occupied, viz. by either *immer* or *auch*; in a closed combination, both particles occupy field II, while field IV is left empty. Standardized residuals are given in brackets.

	open combination	closed combination	total
lexical subject	170 (-5.7)	131 (14.7)	301
pronominal subject	835 (3.3)	18 (-8.7)	853
Total	1,005	149	1,154

We find a highly significant association between lexical subjects and closed combinations on the one hand, and pronominal subjects and open combinations on the other ($\chi^2 = 335.65$; df = 1; $p < 0.0001$), with a strong association overall (Cramér's V = 0.5). All standardized residuals deviate significantly from the expected values (a residual larger than |2| indicates a significant deviation from the expected cell proportion), yet the deviations are especially strong in closed combinations. The slightly weaker deviations in open combinations can be explained by the general tendency for *immer* and *auch* to occupy different fields individually (cf. above).

The observed tendency is easily explained. Pronouns in German have a general left-tendency and usually occur in the left periphery of the middle field (i.e. field II), which is known as "Wackernagel's position" (Lenerz 1993: 117f.). In primary constructions, the constituent occupying this position immediately follows the *W*-word or *W*-phrase. The fact that pronouns compete for this position with *immer* (and certain other elements, e.g. the conjunction *aber* in (18) above) is motivated by information structure: pronouns are typically thematic, i.e. they

express discourse-old, given information, and thus typically occur before rhematic, i.e. discourse-new information (Noel Aziz Hanna 2015: 46). However, it is usually *immer* that gets to occupy this position, since occupying any other position would drastically increase the risk of misinterpretation, whereas pronouns are unproblematic even if they are not the second constituent of the clause (ibid.: 233). Pronominal subjects are thus positively associated with open *immer* (…) *auch* because this combination allows both irrelevance particles to occupy their fields of preference without disturbing the leftward tendency of the pronoun too much.

Conversely, the base position of nominal subjects in German is [Spec, VP] (Lenerz 1993: 118), i.e. the right periphery of the middle field (i.e. field IV). Not only are nominal subjects typically more rhematic than pronouns – thus tending to let the pronouns precede them –, they are obviously also longer and weightier. As a result, the principle of end-weight and the "Law of Increasing Constituents" (Behaghel 1909) become relevant. Given their rhematicity and constituent length, nominal subjects generally prefer the right periphery of the middle field, sometimes forcing *auch* to co-occupy field II with *immer*. The same principles explain the restrictions on the single irrelevance particles *immer* and *auch*: *immer* never follows nominal subjects because these do not compete for Wackernagel's position, whereas *auch* virtually never precedes pronominal subjects, since *auch* does not compete for this position.

The other particle combination, *auch immer*, is less strongly related to the nature of the subject as *immer* (…) *auch*, as seen in Table 4b:

Table 4b: Types of subject and *auch immer* in fields II/IV.

	field II	field IV	total
lexical subject	398 (1.6)	29 (-3.9)	427
pronominal subject	556 (-1.2)	125 (3.1)	681
Total	954	154	1,108

Although the result of the chi-square-test is significant ($\chi^2 = 28.37$; $df = 1$; $p < 0.0001$), the association is rather weak (Cramér's $V = 0.2$), i.e. the nature of the subject is only weakly associated with the positional tendencies of *auch immer*. Since standardized residuals in field II do not deviate significantly from the expected results, the left-leaning tendency of *auch immer* is not influenced significantly by the nature of the subject. When *auch immer* does occupy its dispreferred field IV, however, it tends to do so after pronominal subjects. The underlying reason is, again, the general left-leaning tendency caused by thematicity in pronouns, as argued above.

4.4 Particle combinations between disambiguation and overcharacterisation

With particle combinations, we return to the role of disambiguation as a factor in the distributional patterns seen in our data. There are good reasons, for example, to regard *auch immer* as a less ambiguous substitute for *auch* (cf. Leuschner 2013: 57). As pointed out above, the risk of ambiguity is high when *auch* occupies field II on its own. By contrast, *auch immer* is unambiguously an irrelevance marker, and this in turn explains why *auch immer* shows such a strong preference for II (86.1 %).

(24) a. *Was auch* die Mexikaner anpacken […]
'Whatever the Mexicans tackle' / 'What also the Mexicans tackle'
b. *Was auch immer* die Mexikaner anpacken […]
'Whatever the Mexicans tackle'

In (24a.-b.), repeated from (17) above, the *W*-word is not the subject. (25a.–b.) illustrate the same effect in primary constructions in which the *W*-word is the subject:

(25) a. *was auch passiert* (M01/JUN.44510)
'whatever happens' / 'whatever happens'
b. *Was auch immer passiert* […] (LTB11/JUN.00726)
'Whatever happens'

In (24a.) and (25a.), *auch* could either be an irrelevance particle or a focus particle. In (24b.) no reading of *auch* as a focus particle is possible, and although (24b.) could in principle be read as 'what also always happens', this interpretation is much less plausible than a straightforward irrelevance reading. As mentioned above, *auch immer* is significantly more frequent in subclauses like (25) in which the *W*-word is the subject, and *auch* significantly less. This is likely to be motivated by the fact that *auch immer* is less ambiguous, regardless of position, than *auch*.

Note that we have avoided saying that the addition of *immer* disambiguates *auch*. Although *auch immer* must have arisen as an ad hoc "open" combination of individual particles in the past, our synchronic analysis of it as a single, complex particle suggests that a reanalysis took place at some as yet unspecified stage in history, thenceforth ruling out compositionality. It is therefore more adequate to say that *auch immer* as a unit may take the place of *auch* on its own. Once *auch immer* is used in secondary constructions, where it is very dominant (91.05 %), plenty of opportunities arise for a second reanalysis, this time encompassing the *W*-word. This is how the discourse marker *wie auch immer* 'however', inter alia, must have been created, which however is not part of our sample in this study.

The open combination *immer* (…) *auch*, by contrast, has no such prospects, as shown by its minuscule share of secondary constructions (just 0.37 %, cf. Table 3). In primary constructions, it sometimes functions as a variant of *immer* to which *auch* is added for purposes of disambiguation. This can be useful in those rare instances where *immer* occupies field IV after a pronominal subject. In such cases, (26b.) is more likely to be read as an irrelevance particle than (26a.):

(26) a. *Was er immer* […] *sagt* (RHZ06/MAR.23289)
'Whatever he says' / 'What he always says'
b. *Was man immer auch sagt* (SOZ13/FEB.04565)
'Whatever one says'

Another context in which *immer* can be ambiguous are primary constructions in which the *W*-word is the subject:

(27) a. *Was immer passiert* […] (A99/FEB.11037)
'Whatever happens' / 'What always happens'
b. *Was immer auch passiert* […] (O95/JAN.07794)
'Whatever happens'

Immer auch is relatively rare in such clauses (6.87 %, cf. Table 2b) compared with *auch immer* (13.91 %), yet together they barely dent the dominance of *immer* on its own (78.37 %). *Immer auch* is even less frequent in field II (1.64 %, cf. Table 2a) in clauses with a separate, clause-internal subject. This is unsurprising given that *immer* in field II (67.05 %) cannot normally be read as anything other than an irrelevance particle and is therefore not in need of disambiguation, whether in field II or IV. Yet another matter is *immer … auch* (i.e. straddling the clause-internal subject, 11.09 %): here both particles are in their preferred positions where neither requires disambiguation. Such cases therefore represent overcharacterisation: more irrelevance markers are used than are functionally required.

A closer look at the data brings to light more complex marking strategies which may constitute either disambiguation or overcharacterisation. For example, when *auch* occurs in field II, there is a statistically significant tendency for the finite verb to be a form of the modal verb *mögen* 'may', as in (9a.), repeated here for convenience as (28):

(28) *Was auch* die Gründe sein <u>mögen</u>, nur jammern […] hilft auch nicht weiter. (A01/OKT.32079)
'Whatever the reasons may be, just complaining won't help either.'

Given the non-specific semantics of free-choice quantification and concessive conditionality, modalisation is a well-motivated strategy to support the irrelevance

reading of the clause and thus also of any ambiguous particle. Not surprisingly, 27.27 % of clauses with irrelevance *auch* in field II as in (28) contain a form of *mögen* (n = 6 out of 22, type a in Table 2a) vs. only 6.03 % with irrelevance *auch* in field IV, where *auch* is much less ambiguous (n = 39 out of 647, type f; two-tailed two-proportions Z-test: p < 0.0001). It is therefore safe to describe *mögen* combined with *auch* as a strategy of disambiguation, with some minor spillover leading to overcharacterisation. Compare this with *immer*: *mögen* occurs in 30.77 % of clauses where irrelevance *immer* is in field IV and therefore ambiguous (n= 12 out of 39, type i in Table 2a), but the difference is not significant (p = 0.13), as *mögen* also occurs in 23.11 % of clauses where *immer* is in field II and therefore unambiguous (n = 1,404 out of 6,075, type d). In combination with *immer*, *mögen* therefore tends to represent overcharacterisation, but this does not exclude it from serving genuine disambiguation on occasion.

A clear case of overcharacterisation arises when *mögen* is used in the subjunctive as in (28):

(29) […] - *wer immer* das sein möge - […] (P97/JAN.03698)
'Whoever that may be'

However, this type of overcharacterisation is rare (n = 47; 0.26 % of all primary constructions). A different type of overcharacterisation is seen in (30), where *egal* 'no matter' is added in front of the subclause as a lexical marker of free-choice quantification:

(30) Egal, *was* sie *auch* tun (T06/DEZ.00330)
'No matter what (lit. *whatever*) they do'

Whereas we have double modalisation by lexical means and subjunctive morphology in (29), (30) is best characterised as a contamination of two distinct subtypes of UCCs: one in which the quantification is expressed clause-internally by means of *auch* and/or *immer*, and one in which clause-external adverbs like *egal*, *gleichgültig* ('indifferent') etc. precede the *W*-word in combinations which arose historically from elliptical matrix clauses similar to English *(it is) no matter WH* (Leuschner 2006: Ch. 6). It is generally assumed that clause-internal and clause-external irrelevance marking are in complementary distribution (Breindl 2014:980), with very occasional contaminations of *egal W* and *W … auch*, i.e. *egal W … auch* (Leuschner 2006: 41). Although our data confirm that such contaminations are rare (n = 97; 0.53 % of all primary constructions), the pattern *egal W … auch* is nonetheless more frequent than previously assumed: 8.16 % of all primary constructions with *auch* are contaminations with *egal* and similar adverbs, compared to < 1 % for other particles.

5 Conclusion

The present study has documented and analysed the distributional patterns of the particles *immer* and *auch* in *W*-initial, primary irrelevance clauses and elliptically reduced, secondary constructions, thereby partially replicating Leuschner's (2000) study on the basis of a vastly increased dataset from the *DeReKo* corpus. Our data confirm the complementary positional tendencies of *immer* and *auch*, with *immer* showing a near-exclusive preference for strict adjacency to the *W*-word and *auch* displaying a strong tendency to occupy the right periphery of the middle field. The functionally motivated positional preferences of the individual particles and of their combinations, the difficulties encountered by *immer* vis-à-vis the *W*-word despite its preference for strict adjacency, and the distinct behaviours of *auch immer* as a 'closed' particle combination and *immer* (…) *auch* as an 'open' combination – all reinforce the impression of an emergent subsystem whose evolutionary tendencies are probabilistic in nature rather than deterministic. This is furthermore suggested by occasional spillover into different forms of overcharacterisation on the one hand and a double reanalysis on the other hand which first created *auch immer* and then incorporated it, in secondary constructions, into individual W + *auch immer* combinations like the discourse marker *wie auch immer*.

Follow-up research could expand these findings in several directions. One path follows naturally from the fact that our sample covers only the core *W*-words *was* and *wer*; a true replication of Leuschner (2000) would also refer to *wie* 'how', *warum/weshalb/weswegen/wieso* 'why', *wann* 'when' and *wo* 'where', a mammoth task in view of the high frequency of these words and the need to include *wo*-compounds like *woher/wohin* 'where from/to', *womit* 'where-with, i.e. with which/what', *wogegen* 'where-against, i.e. against which/what' etc. A second path for future research leads to the more systematic inclusion of *egal W*-type markers, linking the dynamism of irrelevance marking at subclause level to the grammaticalisation of entire concessive conditional sentence constructions (Leuschner 2006). A third path would refer specifically to oral data, opening a window on the use and variation of irrelevance marking in spoken German, with a likely focus on the grammaticalisation of secondary constructions. Finally, a promising future perspective on irrelevance marking is crosslinguistic, i.e. typological or intragenetic, contrasting e.g. the synchronic variation and diachronic evolution of German irrelevance marking with the corresponding systems in Dutch (*W* (...) (*dan*) *ook* '*WH* (...) (then) also, i.e. *WH-ever*', *om het even W* 'no matter *WH*') and English. Comparison with these closely related languages is likely to highlight yet again the complex nature of irrelevance marking in German and to conclude, inter alia, that the conceivable grammaticalisation of *WH-ever*-like subordinators from *W*-word + *immer* combinations is likely to remain a protracted building-site in German for the foreseeable future.

References

Behaghel, Otto. 1909. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanische Forschungen* 25: 110–142.

Bossuyt, Tom. 2016. Zur Distribution von Irrelevanzpartikeln in *was immer/auch*-Konstruktionen. Positionelle und kombinatorische Varianz im Deutschen Referenzkorpus. *Germanistische Mitteilungen* 42.1: 45–70.

Breindl, Eva. 2014. Irrelevanzkonditionale Konnektoren. In Eva Breindl, Anna Volodina and Ulrich H. Waßner, *Handbuch der deutschen Konnektoren 2: Semantik der deutschen Satzverknüpfer*, 964–1009. Berlin/New York: de Gruyter.

d'Avis, Franz. 2016. Satztyp als Konstruktion – Diskussion am Beispiel 'Konzessive Konditionalgefüge'. In Rita Finkbeiner and Jörg Meibauer (eds.), *Satztypen und Satzkonstruktionen*, 267–295. Berlin: de Gruyter.

Haspelmath, Martin. 1997. *Indefinite pronouns*. Oxford: Oxford University Press.

Haspelmath, Martin and Ekkehard König. 1998. Concessive conditionals in the languages of Europe. In Johan van der Auwera (ed.), *Adverbial constructions in the languages of Europe*, 563–641. Berlin/New York: de Gruyter.

König, Ekkehard. 1986. Conditionals, concessive conditionals and concessives: areas of contrast, overlap and neutralization. In Elizabeth C. Traugott, Alice ter Meulen, Judy Snitzer Reilly and Charles A. Ferguson (eds.), *On conditionals*, 229–246. Cambridge: Cambridge University Press.

König, Ekkehard and Peter Eisenberg. 1984. Zur Pragmatik von Konzessivsätzen. In Gerhard Stickel (ed.), *Pragmatik in der Grammatik. Jahrbuch 1983 des Instituts für deutsche Sprache*, 313–332. Düsseldorf: Schwann.

König, Ekkehard and Johan van der Auwera. 1988. Clause Integration in German and Dutch Conditionals, Concessive Conditionals, and Concessives. In John Haiman and Sandra A. Thompson (eds.), *Clause Combining in Grammar and Discourse*, 101–133. Amsterdam/Philadelphia: Benjamins.

Kupietz, Marc and Harald Lüngen (2014): Recent developments in DeReKo. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 2378-2385. Reykjavik: ELRA. http://www.lrec-conf.org/proceedings/lrec2014/pdf/842_Paper.pdf (25.02.2017).

Kupietz, Marc, Cyril Belica, Holger Keibel and Andreas Witt (2010): The German Reference Corpus DeReKo: a primordial sample for linguistic research. In Nicoletta Calzolari et al. (eds.), *Proceedings of the 7th conference on International Language Resources and Evaluation (LREC 2010)*, 1848-1854. Valletta: ELRA. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf (25.02.2017).

Lehmann, Christian. 1995. *Thoughts on grammaticalization*. Revised and expanded edition. München: Lincom Europa.

Lenerz, Jürgen. 1993. Zu Syntax und Semantik deutscher Personalpronomina. In Marga Reis (ed.), *Wortstellung und Informationsstruktur*, 117–154. Tübingen: Niemeyer.

Leuschner, Torsten. 1996. *Ever* and universal quantifiers of time: observations from some Germanic languages. *Language Sciences* 18: 496–484.

Leuschner, Torsten. 2000. '…, *wo immer* es mir begegnet, … – *wo* es *auch* sei.' Zur Distribution von 'Irrelevanzpartikeln' in Nebensätzen mit *W auch/immer*. *Deutsche Sprache* 28: 342–356.

Leuschner, Torsten 2001. Nebensatzkonnektoren des Typs ‚W-Wort + Partikel(n)' (Deutsch *wer auch immer* usw.) im Germanischen. Eine intragenetische Typologie aus areallinguistischer Sicht. *Studia Germanica Gandensia* 2001.2: 3–26.

Leuschner, Torsten. 2005. Nonspecific free relatives and (anti)grammaticalization in English and German. *Folia Linguistica Historica* 25: 45–69.

Leuschner, Torsten. 2006. *Hypotaxis as building-site: the emergence and grammaticalization of concessive conditionals in English, German and Dutch*. Munich: Lincom.

Leuschner, Torsten. 2013. Was Partikeln wohl (auch immer) mit Indifferenz zu tun haben. Funktionale und linguistikdidaktische Perspektiven. *Germanistische Mitteilungen* 39.1: 37–62.

Lühr, Rosemarie. 1998. Verallgemeinernde Relativsätze im Althochdeutschen. In Karin Donhauser and Ludwig Eichinger (eds.), *Deutsche Grammatik: Thema in Variationen. Festschrift für Hans-Werner Eroms zum 60. Geburtstag*, 263–281. Heidelberg: Winter.

Neggers, Joseph and Hee Sik Kim. 1998. *Basic Posets*. Singapore, New Jersey, London, Hongkong: World Scientific.

Noel Aziz Hanna, Patrizia. 2015. *Wackernagels Gesetz im Deutschen. Zur Interaktion von Syntax, Phonologie und Informationsstruktur*. Berlin, Boston: de Gruyter.

Nübling, Damaris. 2005. Von *in die* über *in'n* und *ins* bis *im*. Die Klitisierung von Präposition und Artikel als 'Grammatikalisierungsbaustelle'. In Torsten Leuschner, Tanja Mortelmans and Sarah De Groodt (eds.), *Grammatikalisierung im Deutschen*, 105–131. Berlin: de Gruyter.

Overstreet, Maryann. 1999. *Whales, candlelight, and stuff like that: general extenders in English discourse*. New York: Oxford University Press.

Thieroff, Rolf. 2011. *Wer* und *was*. *Germanistische Mitteilungen* 37.1: 47–64.

Thurmair, Maria. 1989. *Modalpartikeln und ihre Kombinationen*. Tübingen: Niemeyer.

Waßner, Ulrich H. 2006. Zur Relevanz von und zur Irrelevanz bei Irrelevanzkonditionalen. In Eva Breindl, Lutz Gunkel and Bruno Strecker (eds.), *Grammatische Untersuchungen. Analysen und Reflexionen. Gisela Zifonun zum 60. Geburtstag*, 381–399. Tübingen: Narr.

Wöllstein, Angelika. 2014. *Topologisches Satzmodell*. 2nd ed. Heidelberg: Winter.

Zifonun, Gisela, Ludger Hoffmann and Bruno Strecker. 1997. *Grammatik der deutschen Sprache*. Berlin, New York: de Gruyter.

Jörg Didakowski, Nadja Radtke

Deutsche Stützverbgefüge in Referenz- und Spezialkorpora: Vergleichsstudien mit dem DWDS-Wortprofil

Abstract The paper deals with the use of so-called empty verb constructions in different text types. It reports on relevant comparative studies carried out in the Digital Dictionary of the German Language (Digitales Wörterbuch der deutschen Sprache, DWDS) and the DWDS word profile. The latter makes available syntactic co-occurrences, which can be used to search potential empty verb constructions in large corpora without having to resort to manual search routines. The studies compare the occurrence and productivity of selected empty verb constructions in a newspaper corpus, a blog corpus and a corpus which is balanced for text types, making use of the association measures provided by the DWDS word profile.

Keywords Deutsche Stützverbgefüge, Textkorpora, Textsortenbereiche, syntaktische Kookkurrenzen, Assoziationsmaße

1 Einleitung

Dieser Beitrag beschäftigt sich mit Stützverbgefügen (SVG) des Deutschen, also mit solchen Konstruktionen wie z. B. *zum Ausdruck bringen*, *eine Änderung erfahren* oder *Kritik üben*, die aus einem prädikativen Nomen und einem semantisch blassen Stützverb gebildet werden. Diese sind von anderen Konstruktionen wie z. B. *zum Flughafen bringen* oder *auf den Hund bringen*, die ebenfalls eine verbale und eine nominale Komponente beinhalten, abzugrenzen.

Verschiedenste Studien beschäftigen sich mit den SVG und ihrer Verwendung in unterschiedlichen Textsortenbereichen. Die zugrunde liegenden Daten werden dort überwiegend manuell erhoben und ausgewertet. Auch wir wenden uns den SVG im Textzusammenhang zu und ziehen ausgewählte Korpora – die Korpora des Digitalen Wörterbuchs der deutschen Sprache (DWDS) – heran. Die Daten für unsere Studien ermitteln wir dabei mithilfe des DWDS-Wortprofils.

Dieses stellt grammatische Kookkurrenzen bereit, die auf Grundlage großer Textkorpora automatisch ermittelt und statistisch bewertet sind (vgl. a. Didakowski/Geyken 2013). Auf diese Weise kann eine überwiegend manuelle Erhebung innerhalb der Textkorpora vermieden werden. Ferner werden unsere Studien dadurch aus praktischer Sicht überhaupt erst möglich. So wird hier auch eine Vorgehensweise vorgestellt, wie bei solchen Studien, die auf großen Mengen von Sprachdaten basieren, mit angemessenem Aufwand vorgegangen werden kann.

In unseren Studien ermitteln wir zunächst mithilfe des DWDS-Wortprofils von den ausgewählten Stützverben ausgehend potenzielle prädikative Nomina und klassifizieren im Weiteren die entsprechenden Verbindungen nach SVG und Nicht-SVG. Daraufhin widmen wir uns in einer ersten Vergleichsstudie dem Vorkommen der SVG in unterschiedlichen Textsortenbereichen. In einer zweiten Studie geht es darum, wie man die Produktivität der Stützverben und die damit verbundene Vielfältigkeit der SVG nachverfolgen kann. Abschließend beschäftigen wir uns in einer letzten Vergleichsstudie mit dem Verhalten verschiedener Assoziationsmaße in den unterschiedlichen Textkorpora.

2 Stützverbgefüge des Deutschen: Terminologie und Gegenstand

Seit langem wecken SVG das Interesse der Forschung. Bereits Daniels (1963) beschäftigt sich mit nominalen Umschreibungen, unter denen die SVG einzuordnen sind, und ihrer Rolle in der Sprache. Er weist in seiner Arbeit darauf hin, dass nominale Umschreibungen „eine sehr alte sprachliche Erscheinung" darstellen (vgl. Daniels 1963: 10f.), dass ihre überwiegend negative Beurteilung – sie werden u. a. als „aufgeblähte Wendungen", „sprachliche Wassersuppen" oder auch als „Fertigware" bezeichnet (vgl. Daniels 1963: 9f.) – zu überdenken ist und dass ihre wichtigen Leistungen, denen er sich dann in seiner Arbeit zuwendet, bei der Kritik an diesen mitberücksichtigt werden sollten. Zu einem ähnlichen Zeitpunkt gehen von Polenz (1963), Engelen (1968), Heringer (1968) und Klein (1968) ebenfalls auf ein System von Konstruktionen ein, die sie als Funktionsverbgefüge (FVG) bezeichnen, indem sie diese u. a. in Bezug auf Kausativität und Inchoativität beschreiben und von anderen Konstruktionen, die ebenfalls aus einer nominalen und einer verbalen Komponente bestehen, abgrenzen. Nach einer darauf folgenden Reihe vielfältiger Arbeiten zu den FVG[1] wendet sich von Polenz (1987) erneut der Beschreibung der FVG zu und betont, eine begriffliche und terminologische Festlegung sei „nützlich und notwendig" (vgl. von Polenz

1 Vgl. u.a. Schmidt 1968, Herrlitz 1973, Persson 1975, Bahr 1977, Guttmacher 1980, Pape-Müller 1980 und Yuan 1987.

1987: 169). Er unterscheidet demnach verschiedene FVG (kausative wie z. B. *in Bewegung bringen*, inchoative wie z. B. *in Kontakt treten*, durative wie z. B. *im Kontakt bleiben* und passivische wie z. B. *Anerkennung finden*) und führt im Weiteren Nominalisierungsverbgefüge (NVG) als heterogenen Bereich ein, dem einerseits die FVG und andererseits weitere Konstruktionen mit einem inhaltsleeren Verb wie z. B. *einen Besuch machen/abstatten*, (*eine*) *Antwort geben/erteilen* und *Verzicht leisten* angehören (vgl. von Polenz 1987: 169ff.). In darauf folgenden Arbeiten sowie in zahlreichen Grammatiken beachten die Autoren überwiegend die von Polenz (1987) geprägte begriffliche Bestimmung des Gegenstandes und die von ihm festgelegten Bezeichnungen, gehen jedoch in ihren eigenen Betrachtungen damit sehr unterschiedlich um. So findet sich z. B. in der *Grammatik der deutschen Sprache* von Zifonun/Hoffmann/Strecker (1997: 1066ff.) sowie in der *Deutschen Grammatik* von Hoffmann (2016: 262ff.) die begriffliche und terminologische Festlegung nach von Polenz (1987) wieder, wobei diese von Hoffmann (2016) nach Storrer (vgl. 2006: 277f.) erweitert wird, indem NVG, die keine FVG sind, als Streckverbgefüge bezeichnet werden. So (1991) folgt ebenfalls der terminologischen Festlegung von von Polenz (1987) und betrachtet in seiner sprachhistorisch angelegten Untersuchung sowohl FVG als auch NVG. Nicht selten wird jedoch der Gegenstandsbereich ausschließlich auf die FVG eingeschränkt. So behandelt Eisenberg (2013: 305ff.) in seinem *Grundriss der deutschen Grammatik* ausschließlich die FVG. Auch Tao (1997) untersucht z. B. in seiner Studie zum Mittelhochdeutschen nur die FVG. Häufig verstehen Autoren aber FVG auch in einem weiteren Sinn und bezeichnen sowohl FVG als auch NVG als FVG. So werden z. B. in der *Deutschen Grammatik* von Helbig/Buscha (2001: 68ff.) sowie in der *DUDEN-Grammatik* (2016: 425ff.) die entsprechenden Konstruktionen als FVG eingeführt und beschrieben. Auch Kamber (2008) verwendet in seiner Untersuchung zu den nominalen Prädikaten des Deutschen einen weiten Begriff und führt die entsprechenden Konstruktionen als FVG auf. Nicht selten werden aber in Studien auch die differenzierenden Bezeichnungen (NVG und FVG) übernommen, jedoch von einzelnen Autoren abweichend bzw. unterschiedlich begrifflich bestimmt und ggf. erweitert. So bezeichnet Ahmed (vgl. 2000: 3 und 29) in seiner Untersuchung zur Abgrenzungsproblematik der FVG gegenüber verwandten Konstruktionen im Deutschen seinen gesamten Untersuchungsbereich als NVG bzw. als prädikative Verbgefüge und versteht unter FVG zentral Konstruktionen wie z. B. *zum Ausdruck bringen* oder *der Meinung sein*, ordnet ihnen peripher aber auch solche wie z. B. *Bezug nehmen* oder *Anerkennung finden* zu und führt abschließend Konstruktionen wie z. B. *Übereinstimmung besteht* oder *einem Irrtum unterliegen* als Streckformen auf.

Die obigen Ausführungen zeigen, dass bis heute weder eine terminologische noch eine begriffliche Einigkeit sowohl bei der Bezeichnung als auch bei der Bestimmung der oben angegebenen Konstruktionen herrscht. Diese stellen

jedoch ein wichtiges Sprachphänomen dar, bei dem es sich um eine sinnvolle Kategorie und einen relevanten Untersuchungsgegenstand handelt.[2]

Bei der Bezeichnung des Untersuchungsgegenstandes orientieren wir uns an den Ausführungen von Langer (2009), der in seiner Arbeit unterschiedliche Termini zur Bezeichnung von Konstruktionen aufführt und diese diskutiert.[3] Daraufhin bezeichnen wir die von uns zu untersuchenden Konstruktionen wie z. B. *Kritik üben* als Stützverbgefüge[4], die nominale Komponente der Konstruktionen als prädikatives Nomen (PN) und ihre verbale Komponente, also das semantisch blasse Verb, als Stützverb (SV). Das Verb, mit dem die jeweilige Konstruktion oft paraphrasiert werden kann und das der Bedeutung der Konstruktion zugrunde liegt, wie hier z. B. *kritisieren*, nennen wir Basisverb. Die Bezeichnung *SVG* deckt den gesamten Bereich der in unseren Studien zu untersuchenden Konstruktionen ab. Sie ist verständlich und hebt hervor, dass das SV die gesamte Konstruktion (das Gefüge) stützt. SVG ist aus unserer Sicht als Bezeichnung nicht vorbelastet und wird nicht – wie es bei FVG (als Bezeichnung einer Teilmenge dieser Konstruktionen) der Fall ist – mit einer bestimmten Funktion[5] in Verbindung gebracht. Wir halten es ebenfalls nicht für sinnvoll, die Bezeichnung *FVG* aufgrund ihrer Geläufigkeit und ihres häufigen Vorkommens auch für den gesamten Bereich der angezielten Konstruktionen zu verwenden.[6] NVG als Bezeichnung für den gesamten Bereich der von uns zu untersuchenden Konstruktionen ziehen wir aufgrund der Fokussierung und Einschränkung auf Nominalisierung ebenfalls

2 Van Pottelberge (2001) unterzieht in seiner Arbeit die Kategorie der – in seiner Terminologie – verbonominalen Konstruktionen (sie entsprechen den FVG) als grammatischen Gegenstand einer kritischen Betrachtung; Winhart 2005 diskutiert ebenfalls in ihrer Arbeit, ob es gerechtfertigt ist, FVG als grammatischen Gegenstand zu betrachten und zu beschreiben.

3 Langer (2009: 41ff. und 68ff.) geht dabei auf die aus der deutschsprachigen Forschung stammende Bezeichnung *Funktionsverbgefüge* und den damit verbundenen Ausdruck *Nominalisierungsverbgefüge* sowie die in der französischen Forschung geprägte Bezeichnung *constructions à verbe support* und die darauf folgende ins Englische übertragene Bezeichnung *support verb construction* ein, die er als Stützverbkonstruktion (STVK) übersetzt.

4 Um mögliche Missverständnisse zu vermeiden, entscheiden wir uns an dieser Stelle bewusst gegen die Bezeichnung *Stützverbkonstruktion.*

5 Siehe z. B. die für die Funktionsverben charakteristischen grammatischen Funktionen sowie ihre im Unterschied zu den jeweiligen Vollverben anderen Bedeutungen (wie z. B. die Bezeichnung des Beginns oder der Dauer eines Vorgangs oder einer Handlung: *in Bearbeitung sein* und *in/zur Bearbeitung kommen*) bei Heringer (2001: 109f.) und die semantisch-syntaktische Funktionsverteilung in FVG bei van Pottelberge (2001: 63).

6 So entscheidet sich Langer (2001: 68ff.), sich an Helbig/Buscha orientierend, aufgrund der stärkeren Verbreitung der Bezeichnung für Funktionsverbgefüge; Kamber (2008: 34) wählt für seine Untersuchung ebenfalls diese gängige Bezeichnung.

nicht in Betracht. Abgesehen davon konnte sich diese Bezeichnung bis heute nicht durchsetzen (vgl. a. Langer 2001: 68).

Bei der Bestimmung des Untersuchungsgegenstandes lehnen wir uns an die Überlegungen von Seifert (2004: 53ff.) an und grenzen die SVG von anderen Konstruktionen ab, die ebenfalls eine verbale und eine nominale Komponente beinhalten. Dies sind einerseits freie Konstruktionen wie z. B. *zum Flughafen bringen* oder *Adresse erfahren* und andererseits Idiome wie z. B. *auf den Hund bringen* oder *Schulterschluss üben*. Für unsere Studien halten wir folgende Typen der SVG[7] fest:

1. SVG mit einem PN als Präpositionalgruppe

1a.
Bei dem PN handelt es sich um ein Abstraktum (Nomen actionis). Dieses kann deverbal wie in *zum Ausdruck bringen* (als Basisverb *ausdrücken*) oder deadjektivisch wie hier *verlegen* in *in Verlegenheit bringen* gebildet sein; das jeweilige SVG ist leicht zu paraphrasieren.

(1) Die deutschen Wörter sind so philosophisch und können Ideen *zum Ausdruck bringen*.
(Die Zeit, 19.11.2012, Nr. 47)[8]

(2) Er ist ein ordentlicher junger Steuermann, und ich habe ihn *in Verlegenheit gebracht*, als ich seine Einladung annahm.
(Andersch, Alfred: Sansibar oder der letzte Grund, Olten: Walter 1957 [1957], S. 104)

1b.
Bei dem PN handelt es sich um ein Abstraktum. Dieses ist unikal oder synchron als Nomen actionis wie in *in Betracht kommen* nicht analysierbar. Es kann in übertragener Bedeutung wie in *in Gang bringen* vorkommen oder es kann ein Fremdlexem wie hier *Bredouille* (‚Verlegenheit') in *in die Bredouille bringen* sein.

7 Bei seinen Ausführungen bezeichnet Seifert (2004: 69ff.) den überwiegenden Teil der einschlägigen Konstruktionstypen als FVG und führt anschließend die NVG ein – es handelt sich dabei um die Konstruktionen mit Abstraktum im Nominativ wie z. B. *die Zahlung erfolgt*. In unseren Studien betrachten wir diese nicht als SVG.

8 In den Beispielbelegen haben wir das SVG je kursiv gesetzt (das jeweilige SV und das jeweilige PN mit der ggf. dazugehörigen Präposition); die vorfindliche Rechtschreibung wurde beibehalten.

(3) Eine Ergänzung des Staatsvertrages *komme* nicht *in Betracht.* (Nr. 302: Besprechung Seiters mit den Chefs der Staats- und Senatskanzleien vom 7. Juni 1990. In: Deutsche Einheit, Berlin: Directmedia Publ. 2000 [1990], S. 3487)

(4) Aber wir möchten den Prozess *in Gang bringen*, daran arbeiten wir hier. (Die Zeit, 24.05.2007, Nr. 22)

(5) Das würde die RTL-Macher *in* die *Bredouille bringen*, schließlich muss die lukrative Pausenwerbung untergebracht werden. (Die Zeit, 02.03.2012 [online])

1c.
Bei dem PN handelt es sich wie bei 1b. um ein Abstraktum, dieses ist jedoch mit der dazugehörigen Präposition wie in *zustande bringen* verschmolzen. Das jeweilige SVG kann ausschließlich diachron analysiert werden.

(6) Dann rollte ein kosmisches Rülpsen über den Ozean, das kein Riesenbollogg *zustande gebracht* hätte. (Moers, Walter: Die 13 1/2 Leben des Käpt'n Blaubär, Frankfurt a. M.: Eichborn 1999, S. 672)

2. SVG mit einem PN im Akkusativ

2a.
Bei dem PN handelt es sich um ein Abstraktum (Nomen actionis). Dieses kann deverbal wie in *Änderung erfahren* (als Basisverb *ändern*) und wie in *Kritik üben* (als Basisverb *kritisieren*) oder deadjektivisch wie hier *aufmerksam* in *Aufmerksamkeit erfahren* gebildet sein; das jeweilige SVG ist leicht zu paraphrasieren.

(7) Insbesondere die Widerrufsrechte werden sowohl inhaltliche als auch gesetzessystematische *Änderungen erfahren.* (http://malekbarudi.info/2013/06/16/bundestag-verab schiedung-gesetzentwurf-eu-verbraucherrechterichtlinie/ 16.06.2013)

(8) Der Trend Change Kommunikation *erfährt* durch die aktuelle wirtschaftliche Situation immer mehr *Aufmerksamkeit.* (http://changekommunikation.wordpress.com/2009/12/03/ change-communications-conference-in-london/ 03.12.2009)

(9) Man kann mitlesen, via Smartphone nachfragen, sich von extern beteiligen, *Kritik üben*, loben oder Missfallen ausdrücken.
(http://werkstatt.bpb.de/2011/12/twitter-eine-kulturkritik/ 01.12.2011)

2b.
Bei dem PN handelt es sich um ein Abstraktum. Dieses ist unikal oder synchron als Nomen actionis wie in *Maßnahmen treffen* nicht analysierbar. Es kann in übertragener Bedeutung wie in *Wendung nehmen* vorkommen oder es kann ein Fremdlexem wie hier *Abstinenz* (‚Enthaltsamkeit') in *Abstinenz üben* sein.

(10) Ich müsste sonst *Maßnahmen treffen.*
(http://imy#-schwamm-drueber/ 24.10.2013)

(11) Genau so sollte man bei diesem Album auch verfahren, weil so mancher Song eine überraschende *Wendung nimmt* und insgesamt richtig Spaß macht.
(http://www.musicampus.de/2009/06/ 01.06.2009)

(12) Aber es nervt halt einfach etwas, wenn ihr dauernd so tut, als könne der Staat geschlechterpolitisch *Abstinenz üben.*
(Die Zeit, 16.05.2012, Nr. 21)

3 Textsortenspezifik und Stützverbgefüge

Betrachtet man die zahlreichen Arbeiten zu SVG, stellt man fest, dass sich die Forschung in den letzten Jahrzehnten mit besonderem Interesse der Verwendung der SVG im Textzusammenhang zuwendet. So untersucht z. B. Schmidt (1968) die Streckformen in publizistischen Texten sowie in der Belletristik, Popadić (1971) beschäftigt sich mit den Nominalisierungen des Verbalausdrucks im Zeitungsdeutsch, Guttmacher (1980) stellt in den Mittelpunkt ihrer Betrachtungen FVG in ausgewählten Zeitungen, Zeitschriften sowie in literaturwissenschaftlichen und belletristischen Werken, Köhler (1985) wendet sich den Funktionsverben in Fachtexten zu, Handschack (1989) betrachtet FVG in sprachwissenschaftlichen Texten, Stein (1993) untersucht verbonominale Prädikate in Patentschriften, Seifert (2004) beschäftigt sich mit den FVG und Nominalisierungsverbgefügen in der Gesetzessprache und Storrer (2013) wendet sich in ihren Studien neben der Verwendung der Streckverbgefüge in Belletristik, Gebrauchstexten, Wissenschaft

und Zeitung[9] auch ihrer Verwendung in juristischen Zeitschriften sowie auf den Artikel- und Diskussionsseiten der deutschen Wikipedia zu.

In den oben erwähnten Studien arbeiten die Autoren nicht nur die Unterschiede in der Verwendung der SVG in ausgewählten Textsortenbereichen heraus, sondern beschäftigen sich u. a. auch mit den Motiven ihrer Verwendung. Die Textsortenspezifik, die dadurch hervorgehoben wird, ist für die Beschreibung der SVG in Wörterbüchern, Grammatiken, Lehrwerken für Deutsch als Fremdsprache sowie in Stilratgebern und -lehren von großer Bedeutung, findet jedoch leider – außer in Stilratgebern und -lehren – eher geringe Beachtung.[10] Und auch wenn die Stilkritiker die Textsortenspezifik der SVG in ihren Werken berücksichtigen, gehen sie dabei jedoch überwiegend einseitig vor, indem sie einfach SVG bestimmten Textsortenbereichen zuordnen, auf eine Erläuterung der Motive ihrer Verwendung aber überwiegend verzichten, um dann generell von der Verwendung dieser Konstruktionen abzuraten.[11]

Angesichts dieser immer noch bestehenden Lücke bei der Beschreibung der Stützverbgefüge wenden wir uns in unseren Studien der Textsortenspezifik der SVG zu, indem wir diese in unterschiedlichen größeren Textkorpora betrachten. Unsere Studien basieren auf den Textdaten aus einem Referenzkorpus mit den Textsortenbereichen *Belletristik, Gebrauchsliteratur, Wissenschaft, Zeitung* und *transkribierte Texte gesprochener Sprache* sowie auf den Textdaten aus zwei Spezialkorpora – aus einem Zeitungskorpus und aus einem Blog-Korpus. Die Textdaten aus den Weblogs, die in den einschlägigen Arbeiten zu SVG bis jetzt keine Berücksichtigung gefunden haben und den Normen redigierter Schriftlichkeit nicht unterliegen, sind dabei für uns von besonderem Interesse und besonderer Relevanz.

Es ist uns bewusst, dass der Begriff *Textsorte* unterschiedlich gefasst wird.[12] Wir gehen in unseren Studien zur Textsortenspezifik der SVG von bestimmten Textsortenbereichen[13] aus, die von den jeweiligen Korpora vertreten werden. Das Blog-Korpus deckt beispielweise den Bereich der internetbasierten

9 Es handelt sich dabei um die Textsortenbereiche des DWDS-Kernkorpus.

10 Vgl. u.a. Heine (2006: 139f.), Kamber (2008: 2f.) und Langer (2009: 182).

11 Das gilt nicht nur für ältere, sondern auch für relativ aktuelle Stilratgeber und -lehren: So werden die SVG in der 36. (allerdings seit den 50er Jahren nicht mehr geänderten) Auflage von Ludwig Reiners *Stilfibel. Der sichere Weg zum guten Deutsch* als eine „Form der Hauptwörterei", die für „Langweiler und Kanzleiräte" typisch ist, eingeführt und beschrieben (siehe Reiners 2009: 87). Klaus Mackowiak als Vertreter aktuellerer Ratschläge ordnet die SVG in seinem 2011 erschienenen Ratgeber *Die häufigsten Stilfehler im Deutschen und wie man sie vermeidet* der Amtssprache zu und empfiehlt dementsprechend auch ihre Verwendung nicht (siehe Mackowiak 2011: 71f.).

12 Siehe dazu etwa Adamzik (2008).

13 Vgl. Storrer (2013).

Kommunikation über Weblogs ab, dem u. a. das Personal Weblog als eigene Textsorte zugeordnet werden kann.[14]

4 Ressourcen und Werkzeuge

Im Folgenden gehen wir auf die für unsere Studien relevanten Ressourcen sowie auf das für unsere Studien relevante Werkzeug ein. Bei den Ressourcen handelt es sich um ausgewählte Korpora des DWDS, bei dem verwendeten Werkzeug um das DWDS-Wortprofil, das im Rahmen unserer Studien auf diesen Korpora basiert.

4.1 DWDS-Korpora

Das DWDS-Kernkorpus ist das Hauptreferenzkorpus des Digitalen Wörterbuchs der deutschen Sprache (DWDS). Es besteht aus ca. 100 Millionen Tokens und ist ein ausgewogenes Korpus der deutschen geschriebenen Sprache des 20. Jahrhunderts. Die Texte sind über die gesamte Zeitspanne und über fünf Textsortenbereiche (Belletristik, Gebrauchsliteratur, Wissenschaft, Zeitung und transkribierte Texte gesprochener Sprache) annähernd gleichmäßig verteilt. Das Korpus ist auf der DWDS-Projektseite http://www.dwds.de größtenteils frei zugänglich.

Das Zeit-Korpus enthält alle Artikel und Ausgaben der Wochenzeitung DIE ZEIT, die auf http://www.zeit.de in digitaler Form zur Verfügung stehen. Es ist auf der DWDS-Projektseite frei zugänglich. Zur Zeit unserer Studien reichte das Korpus von 1946 bis 2015 und beinhaltete ca. 400.000 Tokens.

Das Blog-Korpus enthält Beiträge und Kommentare, die auf Blogs veröffentlicht worden sind (dazu vgl. Barbaresi/Würzner 2014). Es soll zukünftig Teil eines Referenzkorpus zur internetbasierten Kommunikation werden. Es besteht aus ca. 100 Millionen Tokens. Die Beiträge und Kommentare stammen aus den Jahren von ca. 2004 bis 2014. Das Blog-Korpus ist auf der DWDS-Projektseite ebenfalls frei zugänglich.

4.2 DWDS-Wortprofil

Das DWDS-Wortprofil ist Teil des Angebots des DWDS und ermöglicht es, ausgehend von einem Abfragewort Kookkurrenzpaare in verschiedenen grammatischen Relationen zu eruieren und nach ihrer reinen Frequenz oder nach einem

14 Vgl. zum Personal Weblog als Textsorte Schildhauer (2014).

anderen Assoziationsmaß zu ordnen.[15] Es werden drei verschiedene Assoziationsmaße unterstützt: 1) die reine Frequenz, 2) das auf dem Dice-Koeffizienten basierende logDice-Maß (vgl. dazu Rychlý 2008) und 3) das auf Mutual-Information basierende MI-log-Freq-Maß (vgl. dazu Kilgarriff/Tugwell 2002). Das Assoziationsmaß wird hierbei in der Regel so gewählt, dass die entsprechende Sortierung für eine bestimmte Aufgabe am besten geeignet ist (vgl. dazu Evert 2008). Die Kookkurrenzpaare werden auf Grundlage ausgewählter Textkorpora des DWDS mithilfe von computerlinguistischen Verfahren vollautomatisch extrahiert. Kilgarriff u. a. (2004) schlagen für die automatische Extraktion grammatischer Kookkurrenzpaare die flache Sketch-Grammar vor, mit der über reguläre Ausdrücke Kookkurrenzpaare für bestimmte grammatische Relationen extrahiert werden können. Ivanova u. a. (2008) zeigen jedoch, dass es für das Deutsche sinnvoll ist, auf eine reichhaltigere linguistische Analyse zurückzugreifen, um zufriedenstellende Ergebnisse zu erzielen. Daher werden beim DWDS-Wortprofil für die Extraktion der Kookkurrenzpaare einerseits die TAGH-Morphologie (vgl. dazu Geyken/Hannefort 2006), eine Finite-State-Morphologie für das Deutsche mit hoher Abdeckung, und andererseits der robuste regelbasierte syntaktische Finite-State-Parser SynCoP (Syntactic Constraint Parser) (vgl. dazu Didakowski 2008a und Didakowski 2008b) verwendet. Des Weiteren werden nachgeschaltete Filter angewendet, um bestimmte systematische Analysefehler des Parsers zu erkennen und die damit verbundenen Analysen auszuschließen. So kann die relativ reichhaltige Morphologie und die freie Wortstellung des Deutschen angemessen behandelt werden. Über die einzelnen Kookkurrenzpartner zu einem Abfragewort kann direkt auf die Korpusbelege zugegriffen werden. Über diese Verlinkungen bleibt die Recherchierbarkeit gewahrt. Das DWDS-Wortprofil ist über die Projektseite des DWDS abfragbar und verwendet ausgewählte Textkorpora des DWDS. Über die Werkzeuge, die dem DWDS-Wortprofil zugrunde liegen, lassen sich aber auch Kookkurrenzdatenbanken für andere Textkorpora oder Textkorpussammlungen erstellen.

5 Studien zu Stützverbgefügen in unterschiedlichen Textkorpora

In den im Abschnitt 3 genannten Arbeiten, die sich mit den SVG im Textzusammenhang beschäftigen, wurden bereits viele Sprachdaten auf Grundlage großer Textkorpora bearbeitet. Bei allen diesen Untersuchungen wurden die

15 Vgl. Geyken/Didakowski/Siebert (2009) für die initiale Version des DWDS-Wortprofils und Didakowski/Geyken (2013) für seine Weiterentwicklung.

Daten allerdings ausschließlich manuell ausgewertet und überwiegend manuell erhoben.[16]

Für unsere Studien nutzen wir für jedes der zu untersuchenden Textkorpora das DWDS-Wortprofil, mithilfe dessen wir von den SV ausgehend potenzielle PN ermitteln. So muss für eine Erhebung der Daten keine aufwändige Textsuche mehr vollzogen werden. Der manuelle Aufwand umfasst hierbei ausschließlich die Sichtung der potenziellen PN, bei der über Verlinkungen auf die einzelnen Texttreffer zurückgegriffen werden kann. In Didakowski/Radtke (2014) wurde bereits gezeigt, dass auf diese Weise die Erhebung des SVG-Bestandes erheblich beschleunigt werden kann. Über das DWDS-Wortprofil kann zudem auf statistische Maße zugegriffen werden, die mit den einzelnen potenziellen PN verknüpft sind. Dies stellt eine weitere grundlegende Quelle für unsere Studien dar.

Im Folgenden gehen wir zunächst auf das Erstellen eines Wortprofils zur Ermittlung und Beschreibung der SVG in unterschiedlichen Textkorpora ein. Daraufhin legen wir fest, welche SV in unsere Studien mit einbezogen werden, und benennen im Weiteren die für unsere Studien relevanten Fragenstellungen.

5.1 Erstellen eines Wortprofils zur Ermittlung und Beschreibung der Stützverbgefüge in unterschiedlichen Textkorpora

Um die Vergleichbarkeit der Ergebnisse zu gewährleisten, wurden die Korpora in Bezug auf ihre Tokenanzahl auf eine annähernd gleiche Größe gebracht. Hierzu wurde das Zeit-Korpus verkleinert, indem zufällig ausgewählte Dokumente aus dem Korpus entfernt wurden. Eine Auflistung der genauen Zahlen zur Dokumentanzahl, Satzanzahl und Tokenanzahl ist in der Tabelle 1 aufgeführt.[17]

Tabelle 1: Zahlen zur Dokumentanzahl, Satzanzahl und Tokenanzahl in den jeweiligen Korpora.

Korpus	Dokumente	Sätze	Tokens
DWDS-Kernkorpus	79.211	5.841.780	121.386.115
Zeit-Korpus	228.986	5.986.103	111.346.945
DWDS-Blogkorpus	249.578	6.398.524	110.003.872

16 So geht Popadić 1971 25 Ausgaben vom November 1965 der Tageszeitung *DIE WELT* durch und sucht nach den zu untersuchenden verbalen Gefügen; Storrer 2013 nutzt in ihren Studien u. a. die Abfragewerkzeuge des DWDS und erhält für ihre Untersuchungen Trefferlisten mit einem gesuchten Verb wie z. B. *bringen* bzw. Trefferlisten zu einem gesuchten Streckverbgefüge wie z. B. *Entscheidung treffen*, die sie dann anschließend manuell auswertet.

17 Die Token- und Satzgrenzen sind für alle Korpora des DWDS maschinell ermittelt. Die Zahlen können demnach abhängig vom Werkzeug und von bestimmten Parametern leicht variieren.

Ausgehend von dieser Korpusbasis wurde ein Wortprofil erstellt, das die Abfrage von Kookkurrenzen innerhalb der einzelnen Korpora ermöglicht. Die Minimalfrequenz für die Kookkurrenzen wurde dabei auf 5 festgesetzt. Anschließend wurden mithilfe des Wortprofils potenzielle SVG ausgehend von ihren SV ermittelt, was bedeutet, dass zu einem potenziellen SV Kookkurrenzpartner abgefragt wurden. Diese Kookkurrenzpartner stellen dann zusammen mit dem Verb potenzielle SVG dar. Um potenzielle PN mit Präpositionalgruppe zu ermitteln, wurde die grammatische Relation „Verb hat Präpositionalgruppe" herangezogen; zur Ermittlung von PN im Akkusativ entsprechend die grammatische Relation „Verb hat Akkusativ-/Dativobjekt". Da das System Schwierigkeiten hat, Kookkurrenzpartner einer der grammatischen Relationen Akkusativobjekt vs. Dativobjekt verlässlich zuzuordnen, sind hier die beiden Relationen zu einer grammatischen Relation zusammengefasst. Die Relation „Verb hat Passivsubjekt" wurde von uns nicht für die Ermittlung von PN im Akkusativ herangezogen, da hier die Menge an Kookkurrenzen und ihre Qualität für unsere Studien nicht ausreichend sind. Die SVG, die ein PN mit einer verschmolzenen Präposition beinhalten[18], wurden aufgrund ihrer Struktur ebenfalls nicht in unsere Studien mit einbezogen. Schließlich: Einige ermittelte Kookkurrenzpartner sind Personalpronomen. Da die entsprechenden Kookkurrenzen für unsere Studien nicht relevant sind, wurden diese von uns aus den Kookkurrenzlisten nachträglich entfernt.

5.2 Auswahl der Stützverben

Bei der Auswahl der SV für unsere Studien gehen wir zunächst von einer anhand ausgewählter Grammatiken (insgesamt 23) erstellten Lemmaliste aus, die 125 Verben beinhaltet, von denen 41 Verben mindestens fünfmal in den jeweiligen Grammatiken vorkommen. Aus diesen 41 Verben wählen wir zehn aus, die für unsere Studien besonders interessant sind. Dabei achten wir darauf, dass die ausgewählten Verben in unterschiedlichen Typen von SVG[19] vorkommen. Die Tabelle 2 zeigt die zehn ausgewählten Verben und die Anzahl der ermittelten Kookkurrenzen pro grammatische Relation (Präpositionalgruppe bzw. Akkusativ-/Dativobjekt) in den jeweiligen Textkorpora.

Von den zehn Verben wählten wir anschließend drei aus, die ebenfalls in unterschiedlichen Typen von SVG vorkommen und einen ergiebigen Vergleich für unsere Studien versprachen. Es handelt sich um die folgenden:

18 Siehe den Typ 1c. bei den von uns festgehaltenen Typen der SVG im Abschnitt 2; die SVG dieses Typs können bei weiteren Studien separat ermittelt und ergänzend zu den anderen Typen betrachtet werden.

19 Siehe dazu die von uns festgehaltenen Typen der SVG im Abschnitt 2.

Tabelle 2: Stützverben und die Anzahl der Kookkurrenzen pro grammatische Relation in den jeweiligen Korpora.

	Stützverb	DWDS-Kernkorpus: Präpositionalgruppe	DWDS-Kernkorpus: Akkusativ-/Dativobjekt	Zeit-Korpus: Präpositionalgruppe	Zeit-Korpus: Akkusativ-/Dativobjekt	DWDS-Blogkorpus: Präpositionalgruppe	DWDS-Blogkorpus: Akkusativ-/Dativobjekt
1	*bringen*	1.011	1.037	1.111	1.039	560	581
2	*erfahren*	107	149	140	69	120	89
3	*finden*	1.423	1.111	1.155	904	1.137	1.249
4	*geben*	1.462	2.536	2.501	3.346	1.957	3.041
5	*halten*	843	619	851	721	467	408
6	*kommen*	2.390	299	2.775	282	1.613	215
7	*leisten*	123	123	92	134	54	76
8	*nehmen*	783	813	570	669	334	513
9	*treffen*	450	205	816	307	231	143
10	*üben*	19	40	25	24	12	10

bringen mit einer Präpositionalgruppe wie im SVG *zum Lachen bringen*

(13) Er hoffe, sein Publikum *zum Lachen gebracht* zu haben.
(Die Zeit, 28.10.2014 [online])

erfahren mit Akkusativ-/Dativobjekt wie im SVG *eine Änderung erfahren*

(14) Diese Behandlung, die bei der Herstellung des Weißweines angewendet wird, muß bei der Rotweinbereitung eine *Änderung erfahren.*
(Kölling, Alfred: Fachbuch für Kellner, Leipzig: Fachbuchverl. VEB 1962 [1956], S. 164)

üben mit Akkusativ-/Dativobjekt wie z. B. im SVG *Kritik üben*

(15) Weil ich es fair finde, daß man seinen Namen nennt, wenn man *Kritik übt.*
(http://lumma.de/2004/03/17/post-an-wagner/ 17.03.2004)

Die Tabellen 3, 4 und 5 geben Auskunft darüber, wie viele Kookkurrenzen zu den ausgewählten Verben ermittelt wurden und welchen Anteil diese Kookkurrenzen an allen Kookkurrenzen in der gleichen grammatischen Relation haben. Gleiches ist für die Anzahl der Texttreffer aufgeführt, die mit den jeweiligen Kookkurrenzen verknüpft sind.

Tabelle 3: Das Verb *bringen* in der grammatischen Relation „Verb mit Präpositionalgruppe".

Korpus	Kookkurrenzen mit dem Verb *bringen*	Anteil an allen Kookkurrenzen	Texttreffer mit dem Verb *bringen*	Anteil an allen Texttreffern
DWDS-Kernkorpus	1.011	0,154375 %	23.082	1,116375 %
Zeit-Korpus	1.111	0,119988 %	28.855	0,993899 %
DWDS-Blogkorpus	560	0,118616 %	12.446	0,838295 %

Tabelle 4: Das Verb *erfahren* in der grammatischen Relation „Verb mit Akkusativ-/Dativobjekt".

Korpus	Kookkurrenzen mit dem Verb *erfahren*	Anteil an allen Kookkurrenzen	Texttreffer mit dem Verb *erfahren*	Anteil an allen Texttreffern
DWDS-Kernkorpus	149	0,040404 %	1.922	0,084001 %
Zeit-Korpus	69	0,014672 %	843	0,037050 %
DWDS-Blogkorpus	89	0,031739 %	961	0,055441 %

Tabelle 5: Das Verb *üben* in der grammatischen Relation „Verb mit Akkusativ-/Dativobjekt".

Korpus	Kookkurrenzen mit dem Verb *üben*	Anteil an allen Kookkurrenzen	Texttreffer mit dem Verb *üben*	Anteil an allen Texttreffern
DWDS-Kernkorpus	40	0,010847 %	756	0,033041 %
Zeit-Korpus	24	0,005103 %	662	0,029095 %
DWDS-Blogkorpus	10	0,003566 %	292	0,016846 %

Hier zeigt sich, dass für das Verb *üben* die wenigsten Kookkurrenzen ermittelt werden konnten und dass auf das DWDS-Blogkorpus insgesamt die wenigsten Kookkurrenzen entfallen. Jedoch ist über die Korpora hinweg das Verhältnis der Kookkurrenzen mit einem der ausgewählten Verben zu allen Kookkurrenzen recht stabil. Gleiches gilt für die Texttreffer. Somit ist gewährleistet, dass die jeweiligen Korpora vergleichbare Ergebnisse liefern und sicher für unsere Studien herangezogen werden können.

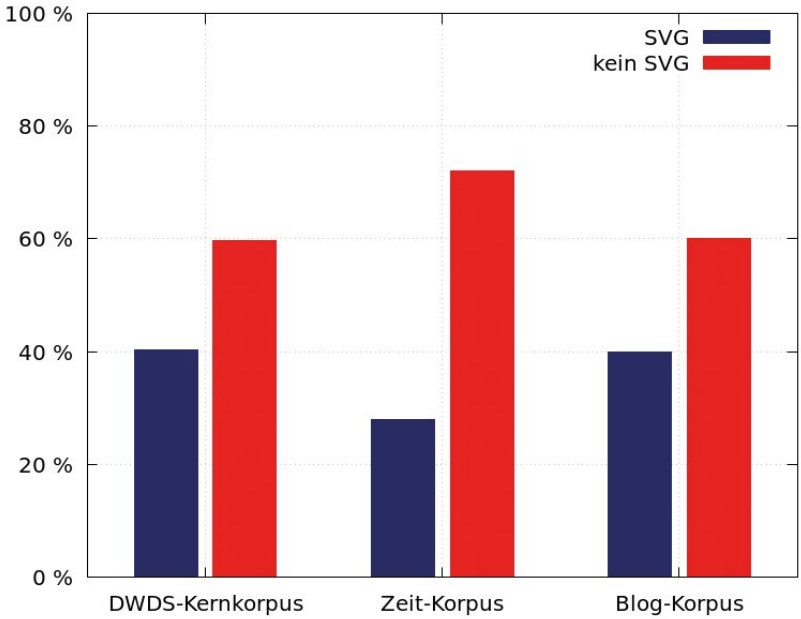


Abbildung 1: Das Vorkommen der SVG in den unterschiedlichen Korpora.

5.3 Stützverbgefüge und Stützverben: Vorkommen, Vielfältigkeit, Produktivität und Assoziationsmaße

Für die folgenden Studien haben wir zunächst die automatisch ermittelten Kookkurrenzen zu den drei ausgewählten Verben *bringen*, *erfahren* und *üben* manuell vollständig gesichtet und nach SVG und Nicht-SVG klassifiziert.

Bei der ersten Vergleichsstudie stand das Vorkommen der SVG in den unterschiedlichen Textsortenbereichen im Zentrum. Wir gingen dabei der Frage nach, wie sich die Anzahl der SVG-Vorkommen in den unterschiedlichen Textkorpora und in Hinblick auf die Ausgewogenheit eines Textkorpus unterscheidet.

Abbildung 1 zeigt das Verhältnis von SVG zu Nicht-SVG in den jeweiligen Korpora. Beim DWDS-Kernkorpus liegt der Anteil der SVG an allen ermittelten Kookkurrenzen bei 40,32 %, der Anteil der Nicht-SVG entsprechend bei 59,68 %. Ähnlich verhält sich das DWDS-Blogkorpus. Hier liegt der Anteil der SVG bei 39,91 %, der der Nicht-SVG bei 60,09%. Beim Zeit-Korpus hingegen liegt der Anteil der SVG bei 27,98%, der der Nicht-SVG bei 72,02 %. Die Zahlen zeigen, dass viele der ermittelten Konstruktionen SVG sind, aber dass ihr Anteil nicht überwiegt. Hierbei sticht das Zeit-Korpus etwas hervor, bei dem das Verhältnis von SVG zu Nicht-SVG bei 0,39 liegt. Bei den anderen beiden Korpora ist hingegen ein etwa doppelt so hoher Wert zu beobachten.

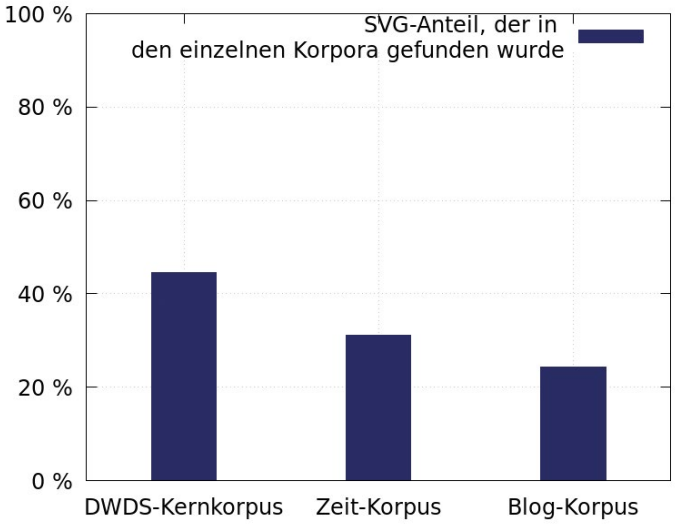


Abbildung 2: Anteil der SVG in den jeweiligen Korpora.

In der zweiten Vergleichsstudie sahen wir uns an, wie die ermittelten SVG für die Verben *bringen*, *erfahren* und *üben* auf die einzelnen Korpora verteilt sind, um die Produktivität der SV und um die damit verbundene Vielfältigkeit der SVG untersuchen zu können. Ein SV kann dann als produktiv eingestuft werden, wenn im Korpus viele verschiedene SVG vorkommen, die mit diesem gebildet werden.

Abbildung 2 zeigt den Anteil der SVG, der auf die jeweiligen Korpora entfällt. Das DWDS-Kernkorpus hat einen Anteil von 44,65 %, das Zeit-Korpus einen von 31,09 % und das DWDS-Blogkorpus einen von 24,26 %. Hier lässt sich also ein gewisses Gefälle beobachten.

Abbildung 3 zeigt demgegenüber den Anteil der SVG, die ausschließlich in einem der Korpora gefunden wurden. 21,48 % der SVG waren ausschließlich im DWDS-Kernkorpus, 7,01% ausschließlich im Zeit-Korpus und 4,98 % ausschließlich im DWDS-Blogkorpus zu finden. Das DWDS-Kernkorpus enthält also mit Abstand die meisten solcher SVG. Das Verhältnis der Anzahl der SVG im DWDS-Kernkorpus zur Anzahl der SVG im DWDS-Zeit-Korpus liegt bei 3,05. Im Vergleich mit dem Blogkorpus liegt das Verhältnis sogar bei 4,30.

Die Abbildungen 2 und 3 verdeutlichen, dass innerhalb eines ausgewogenen Textkorpus die SVG vielfältiger und somit die ausgewählten SV produktiver sind. Aber auch das Zeit-Korpus und das DWDS-Blogkorpus enthalten SVG, die ausschließlich in einem Korpus gefunden wurden. Dies bedeutet, dass bestimmte

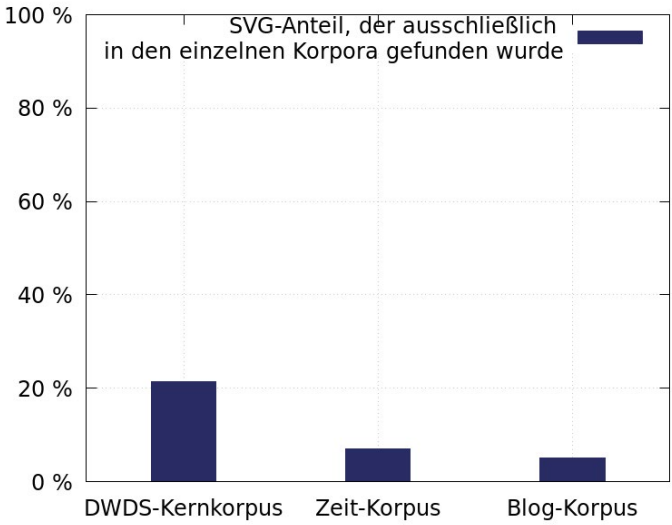


Abbildung 3: Anteil der SVG, die ausschließlich in den jeweiligen Korpora gefunden wurden.

SVG ausschließlich in bestimmten Textsortenbereichen vorkommen und dass beim Heranziehen von Daten aus mehreren Textsortenbereichen – wie hier beim DWDS-Kernkorpus – daher ein vielfältigeres Spektrum an SVG zu erwarten ist.

Im Folgenden werden exemplarisch zwei Beispielbelege für SVG gegeben, die ausschließlich im DWDS-Blogkorpus vorgekommen sind:

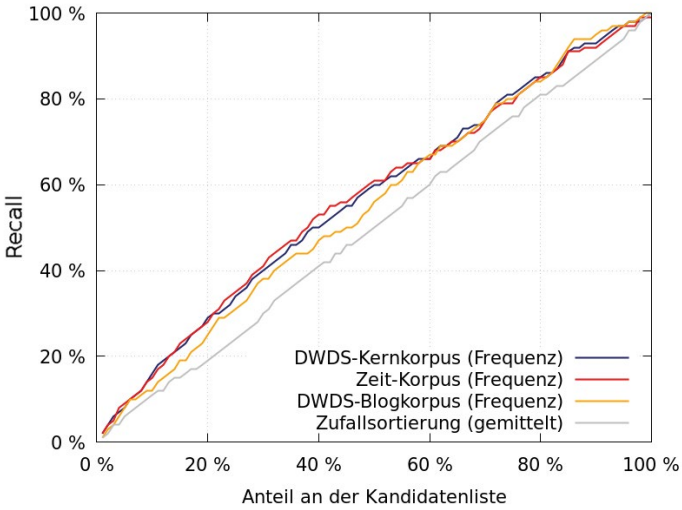
ein Update erfahren

(16) Die Gallerie[!] hat nur ein kleines *Update erfahren*, genauer gesagt waren es sogar nur Sicherheitsupdates. (http://jensman.wordpress.com/2005/12/22/grosses-update/ 22.12.2005)

auf den Blog bringen

(17) Dank Einbindung der Instagram-Bilder werde ich ohnehin regelmäßigen neuen Content *auf* den *Blog bringen*. (http://kaiobi.wordpress.com/2013/07/17/s00c13-vorbei-was-bleibt/ 17.07.2013)

In der dritten Vergleichsstudie wandten wir uns den verschiedenen Sortiermöglichkeiten – der Sortierung nach bestimmten Assoziationsmaßen – der



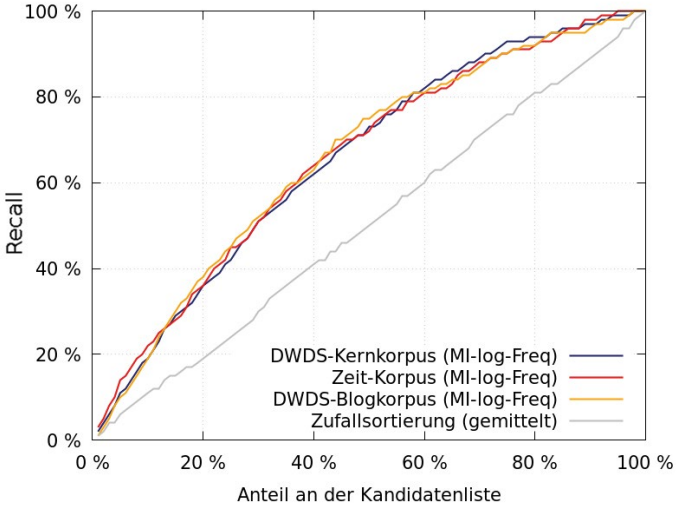


Abbildung 4 (oben): Recall bezüglich der reinen Frequenz;
Abbildung 5 (unten): Recall bezüglich des Assoziationsmaßes MI-log-Freq.

Kookkurrenzlisten innerhalb des DWDS-Wortprofils zu. In Didakowski/Radtke (2014) wurde bereits gezeigt, dass es über ein geeignetes Assoziationsmaß möglich ist, die Kookkurrenzlisten so zu sortieren, dass am Anfang der Listen die Dichte der prädikativen Nomina sehr hoch ist und am Ende nur wenige prädikative Nomina vorkommen. Für diese Aufgabe erweist sich das Assoziationsmaß *MI-log-Freq* als am geeignetsten. Hier soll nun anhand von Recall-Kurven gezeigt werden, inwieweit die Ausgewogenheit eines Korpus und die Textsortenbereiche Einfluss auf die Sortierung mit dem MI-log-Freq-Maß haben.[20] Wir beziehen uns dabei ebenfalls auf die drei ausgewählten Verben *bringen*, *erfahren* und *üben*.

Die Abbildungen 4 und 5 zeigen die Recall-Kurven zu den untersuchten Korpora für die Sortierung nach der reinen Frequenz bzw. für die Sortierung nach dem MI-log-Freq-Maß. Abzulesen ist jeweils, wie viel Prozent einer Kookkurrenzliste durchgesehen werden muss (x-Achse), um einen bestimmten Abdeckungsgrad (y-Achse) zu erreichen. Die Recall-Kurven sind aus den Kookkurrenzlisten der einzelnen Verben (gemittelt) für jedes Korpus berechnet worden. Für die Bewertung der Eignung eines Assoziationsmaßes haben wir zusätzlich eine Zufallssortierung einbezogen, die über die Korpora gemittelt ist (graue Kurven). Die Kurven bei den beiden Abbildungen zeigen, dass eine Sortierung nach der reinen Frequenz oder MI-log-Freq besser ist als eine zufällige Sortierung. Eine Sortierung nach dem MI-log-Freq-Maß liefert deutlich das beste Ergebnis. Wenn man beispielsweise 40 % der Liste betrachtet, sind bei dieser Sortierung bereits 60 % der SVG enthalten. Bei der Sortierung nach der reinen Frequenz wären hingegen nur 50 % enthalten. Interessant ist hierbei, dass beim MI-log-Freq-Maß die Wahl des Korpus keinen Einfluss auf die Sortierung hat. Etwas abgeschwächt ist dies auch bei der reinen Frequenz der Fall. Dies bedeutet, dass die SVG über die verschiedenen Korpora hinweg mit den gleichen syntaktisch-distributionellen Eigenschaften als Konstruktionen auftreten.

6 Zusammenfassung

Im Zentrum unserer Studien standen SVG des Deutschen und ihre Verwendung in unterschiedlichen Textkorpora. Für unsere Studien haben wir ein Wortprofil auf Grundlage eines ausgewogenen Korpus, eines Zeitungskorpus sowie eines Blog-Korpus erstellt. Ausgehend von drei ausgewählten SV wurden anschließend über das erstellte Wortprofil Kookkurrenzlisten ermittelt, die potenzielle SVG enthalten. Nach einer manuellen Durchsicht und Klassifikation der Kookkurrenzlisten nach SVG und Nicht-SVG betrachteten wir das Vorkommen der

20 Vgl. Evert/Heid/Lezius (2000) für die Beurteilung von Assoziationsmaßen.

SVG in den jeweiligen Textkorpora sowie die Produktivität der SV und die Vielfältigkeit der SVG. Wir konnten feststellen, dass die vorkommenden SV in einem ausgewogenen Korpus produktiver und dass dort die vorhandenen SVG vielfältiger sind. Im Kontext der Ermittlung eines Bestandes von SVG wäre diese Beobachtung für die Wahl des Korpus relevant. Dort sollte entsprechend ein ausgewogenes Korpus stets die erste Wahl sein. Ob verschiedene Textsortenbereiche alleine zu unterschiedlicher Produktivität und Vielfalt führen, konnte hingegen für die von uns untersuchten Textsortenbereiche nicht festgestellt werden. Weiterhin ist festzuhalten, dass die Wahl des Korpus keinen Einfluss auf die Eignung der von uns untersuchten Assoziationsmaße hat.

Literaturhinweise und Ressourcen

Adamzik, Kirsten (2008): Textsorten und ihre Beschreibung. In: Janich, Nina (Hg.): Textlinguistik. Tübingen: Gunter Narr, S. 145–175.

Ahmed, Elsayed (2000): Die Nominalisierungsverbgefüge und die prädikativen Verbgefüge. Eine Untersuchung zur Abgrenzungsproblematik der Funktionsverbgefüge gegenüber verwandten Konstruktionen im Deutschen. Neuried: Ars Una.

Bahr, Brigitte Inge (1977): Untersuchungen zu Typen von Funktionsverbgefügen und ihrer Abgrenzung gegen andere Arten der Nominalverbindung. Bonn: Universität Bonn.

Barbaresi, Adrien / Würzner, Kay-Michael (2014): For a fistful of blogs: Discovery and comparative benchmarking of republishable German content. In: Proceedings of KONVENS 2014, Hildesheim, S. 2–10.

Daniels, Karlheinz (1963): Substantivierungstendenzen in der deutschen Gegenwartssprache. Nominaler Ausbau des verbalen Denkkreises. Düsseldorf: Pädagogischer Verlag Schwann.

Didakowski, Jörg (2008a): SynCoP – Combining Syntactic Tagging with Chunking Using Weighted Finite State Transducers. In: Hanneforth, Thomas / Würzner, Kay Michael (Hg.): Finite-State Methods and Natural language Processing, 6th International Workshop, FSMNLP 2007, Potsdam, Germany, September 14–16, revised papers. Potsdam: Potsdam University Press, S. 107–118.

Didakowski, Jörg (2008b): Local Syntactic Tagging of Large Corpora using Weighted Finite State Transducers. In: Storrer, Angelika / Geyken, Alexander / Siebert, Alexander / Würzner, Kay-Michael (Hg.): Text Resources and Lexical Knowledge – Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008. Text, Translation, Computational Processing, Berlin/New York: Mouton de Gruyter, S. 65–78.

Didakowski, Jörg / Geyken, Alexander (2013): From DWDS corpora to a German Word Profile – methodological problems and solutions. In: Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information, 2nd Work Report of the Academic Network "Internet Lexicography" (OPAL-Online publizierte Arbeiten zur Linguistik X/2012). Mannheim: Institut für Deutsche Sprache, S. 43–52.

Didakowski, Jörg / Radtke, Nadja (2014): Nutzung des DWDS-Wortprofils beim Aufbau eines lexikalischen Informationssystems zu deutschen Stützverbgefügen. In: Abel, Andrea / Vettori, Chiara / Ralli, Natascia (Hg.): Proceedings of the XVI EURALEX International Congress: The User in Focus. 15–19 July 2014, Bolzano/Bozen: EURAC research, S. 345–353.

Das Digitale Wörterbuch der deutschen Sprache (DWDS): http://www.dwds.de (21.06.2017).

DUDEN. Band 4. Die Grammatik (2016). Berlin: Dudenverlag.

Eisenberg, Peter (2013): Grundriss der deutschen Grammatik. Band 2: Der Satz. Stuttgart u. a.: J. B. Metzler.

Engelen, Bernhard (1968): Zum System der Funktionsverbgefüge. In: Wirkendes Wort. 18, S. 289–303.

Evert, Stefan (2008): Corpora and collocations. In: Lüdeling, Anke / Kytö, Merja (Hg.): Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter, S. 1212–1248.

Evert, Stefan / Heid, Ulrich / Lezius, Wolfgang (2000): Methoden zum qualitativen Vergleich von Signifikanzmaßen zur Kollokationsidentifikation. In: Schukat-Talamazzini, Ernst Günter / Zühlke, Werner (Hg.): Sprachkommunikation, KONVENS 2000. Berlin/Offenbach: VDE, S. 215–220.

Geyken, Alexander (2007): The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Christiane (Hg.): Idioms and Collocations. Corpus-based Linguistic and Lexicographic Studies. London: Continuum, S. 23–41.

Geyken, Alexander / Didakowski, Jörg / Siebert, Alexander (2009): Generation of word profiles for large German corpora. In: Kawaguchi, Yuji / Minegishi, Makoto / Durand, Jacques (Hg.): Corpus Analysis and Variation in Linguistics. Tokyo University of Foreign Studies, Studies in Linguistics 1, John Benjamins Publishing Company, S. 141–157.

Geyken, Alexander / Hanneforth, Thomas (2006): TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In: Yli-Jyrä, Anssi / Karttunen, Lauri / Karhumäki, Juhani (Hg.): Finite State Methods and Natural Language Processing. Lecture Notes in Computer Science, Bd. 4002, Berlin/Heidelberg: Springer, S. 55–66.

Guttmacher, Karla (1980): Die Stellung der Funktionsverbgefüge im deutschen Verbsystem. Jena: Universität Jena.

Handschack, Joachim (1989): Funktionsverbgefüge in sprachwissenschaftlichen Texten. Eine Analyse unter funktional-kommunikativem Aspekt. Zwickau: Fakultät für Gesellschaftswissenschaften des Wissenschaftlichen Rates der Pädagogischen Hochschule „Ernst Schneller".

Heine, Antje (2006): Funktionsverbgefüge in System, Text und korpusbasierter (Lerner-) Lexikographie. Frankfurt a.M.: Peter Lang.

Helbig, Gerhard / Buscha, Joachim (2001): Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht. Berlin u. a.: Langenscheidt.

Heringer, Hans-Jürgen (1968): Die Opposition von „kommen" und „bringen" als Funktionsverben. Untersuchungen zur grammatischen Wertigkeit und Aktionsart. Düsseldorf: Pädagogischer Verlag Schwann.

Heringer, Hans Jürgen (2001): Lesen lehren lernen: Eine rezeptive Grammatik des Deutschen. Tübingen: Max Niemeyer.

Herrlitz, Wolfgang (1973): Funktionsverbgefüge vom Typ „in Erfahrung bringen". Ein Beitrag zur generativ-transformationellen Grammatik des Deutschen. Tübingen: Max Niemeyer.

Hoffmann, Ludger (2016): Deutsche Grammatik. Grundlagen für Lehrerausbildung, Schule, Deutsch als Zweitsprache und Deutsch als Fremdsprache. Berlin: Erich Schmidt.

Ivanova, Kremena / Heid, Ulrich u. a. (2008): Evaluating a German Sketch Grammar: A Case Study on Noun Phrase Case. In: Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech (Morocco), S. 2101–2107.

Kamber, Alain (2008): Funktionsverbgefüge – empirisch. Eine korpusbasierte Untersuchung zu den nominalen Prädikaten des Deutschen. Tübingen: Max Niemeyer.

Kilgarriff, Adam / Rychly, Pavel / Smrz, Pavel / Tugwell, David (2004): The Sketch Engine. In: Proceedings of the 11th EURALEX International Congress. Lorient (France), S. 105–116.

Kilgarriff, Adam / Tugwell, David (2002): Sketching Words. In Corréard, Marie-Hélène (Hg.): Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins, EURALEX, S. 125–137.

Klein, Wolfgang (1968): Zur Kategorisierung der Funktionsverben. In: Beiträge zur Linguistik und Informationsverarbeitung. 13, S. 7–37.

Köhler, Claus (1985): Verben in deutschsprachigen Fachtexten – Supplementverben (eine Voraussetzung der Nominalität von Fachtextsetzen). In: Fachsprache – Fremdsprache – Muttersprache. 1.

Langer, Stefan (2009): Funktionsverbgefüge und automatische Sprachverarbeitung. München: LINCOM.

Mackowiak, Klaus (2011): Die häufigsten Stilfehler im Deutschen und wie man sie vermeidet. München: C. H. Beck.

Pape-Müller, Sabine (1980): Textfunktionen des Passivs. Untersuchungen zur Verwendung von grammatisch-lexikalischen Passivformen. Tübingen: Max Niemeyer.

Persson, Ingemar (1975): Das System der kausativen Funktionsverbgefüge. Eine semantisch-syntaktische Analyse einiger verwandter Konstruktionen. Lund: CWK Gleerup.

Polenz, Peter von (1963): Funktionsverben im heutigen Deutsch. Sprache in der rationalisierten Welt. Beihefte zur Zeitschrift „Wirkendes Wort" 5. Düsseldorf: Pädagogischer Verlag Schwann.

Polenz, Peter von (1987): Funktionsverben, Funktionsverbgefüge und Verwandtes. Vorschläge zur satzsemantischen Lexikographie. In: Zeitschrift für germanistische Linguistik. 15, S. 169–189.

Popadić, Hanna (1971): Untersuchungen zur Frage der Nominalisierungen des Verbalausdrucks im heutigen Zeitungsdeutsch. Tübingen: Gunter Narr.

Pottelberge, Jeroen van (2001): Verbonominale Konstruktionen, Funktionsverbgefüge. Vom Sinn und Unsinn eines Untersuchungsgegenstandes. Heidelberg: Universitätsverlag C. Winter.

Reiners, Ludwig (2009): Stilfibel. Der sichere Weg zum guten Deutsch. München: Deutscher Taschenbuch Verlag.

Rychlý, Pavel (2008): A lexicographer-friendly association score. In: Sojka, Petr / Horák, Aleš (Hg.): Proceedings of the Second Workshop on Recent Advances in Slavonic Natural Languages Processing. Brno: Masaryk University, S. 6–9.

Schildhauer, Peter (2014): Textsorten im Internet zwischen Wandel und Konstanz. Eine diachrone Untersuchung der Textsorte Personal Weblog. (http://digital.bibliothek.uni-halle.de/hs/content/titleinfo/2007276)

Schmidt, Veronika (1968): Die Streckformen des deutschen Verbums. Substantivisch-verbale Wortverbindungen in publizistischen Texten der Jahre 1948 bis 1967. Halle (Saale): Max Niemeyer.

Seifert, Jan (2004): Funktionsverbgefüge in der deutschen Gesetzessprache (18.–20. Jahrhundert). Hildesheim u. a.: Georg Olms.

So, Man-Seob (1991): Die deutschen Funktionsverbgefüge in ihrer Entwicklung vom 17. Jahrhundert bis zur Gegenwart. Eine sprachhistorische Untersuchung anhand von populärwissenschaftlichen Texten. Trier: Wissenschaftlicher Verlag Trier.

Stein, Achim (1993): Nominalgruppen in Patentschriften. Komposita und prädikative Nominalisierungen im deutsch-französischen Vergleich. Tübingen: Max Niemeyer.

Storrer, Angelika (2006): Zum Status der nominalen Komponenten in Nominalisierungsverbgefügen. In: Breindl, Eva / Gunkel, Lutz / Strecker, Bruno (Hg.): Grammatische Untersuchungen, Analysen und Reflexionen. Festschrift für Gisela Zifonun. Tübingen: Gunter Narr, S. 275–295.

Storrer, Angelika (2013): Variation im deutschen Wortschatz am Beispiel der Streckverbgefüge. In: Reichtum und Armut der deutschen Sprache. Erster Bericht zur Lage der deutschen Sprache. Herausgegeben von der Deutschen Akademie für Sprache und Dichtung und der Union der deutschen Akademien der Wissenschaften. Mit Beitrag von Eichinger, Ludwig / Eisenberg, Peter / Klein, Wolfgang / Storrer, Angelika. Berlin u. a.: de Gruyter, S. 171–209.

Tao, Jingning (1997): Mittelhochdeutsche Funktionsverbgefüge. Materialsammlung, Abgrenzung und Darstellung ausgewählter Aspekte. Tübingen: Max Niemeyer.

Winhart, Heike (2005): Funktionsverbgefüge im Deutschen. Zur Verbindung von Verben und Nominalisierungen. Tübingen: Universität Tübingen. (http://www.dart-europe.eu/full.php?id=89649).

Yuan, Jie (1987): Funktionsverbgefüge im heutigen Deutsch. Eine Analyse und Kontrastierung mit ihren chinesischen Entsprechungen. Heidelberg: Julius Groos.

Zifonun, Gisela / Hoffmann, Ludger / Strecker, Bruno u. a. (1997): Grammatik der deutschen Sprache. 3 Bände. Berlin u. a.: Walter de Gruyter.

Oliver Wicher

Corpus-Driven Lexical Grammar and the Aspect-Modality Interface: The Case of French Past Modal Constructions

Abstract French modal verbs unite temporal, aspectual and modal values in past-tense constructs such as *j'ai voulu faire* vs. *je voulais faire* or *elle a pu rentrer* vs. *elle pouvait rentrer*. The semantics of these *past modal constructions* have been considered a puzzling area, as perfective aspect on root modals forces the complement to take place in the actual world, triggering the so-called 'actuality entailment' effect. The present study analyzes the behaviour of French past modal constructions from a corpus-driven constructional perspective. To this end, the author presents a new reference corpus of French and shows that past-tense choice of French modals can be considered a matter of collostructional preference: perfective and imperfective modals each choose distinct sets of verbal complements forming lexico-grammatical patterns. The results corroborate the actuality entailment hypothesis, and give the opportunity to discuss how the aspect-modality interface in French can be accounted for from a constructional perspective.

Keywords French, aspect, modality, lexical grammar, corpus-driven, collostruction

1 Introduction

The present paper tackles a grammatical phenomenon known for its linguistic intricacy: the past-tense use of French modal verbs. They have been extensively analyzed in formal semantics (Bhatt 1999; Hacquard 2006, 2009; Borgonovo/Cummins 2007; Mari/Martin 2007; Martin 2009; Homer 2011; Laca 2012 among others; cf. Rubio Vallejo 2017 for a pragmatic analysis). However, their descriptive analysis in large-scale corpora has been uncharted territory. It will be demonstrated how a corpus-driven approach can lead to a more precise description of forms, meanings and usage patterns of these *past modal constructions* (PMCs):

(1) a. Quand j'ai voulu passer le conservatoire, j'ai profité d'un déjeuner pour lui dire ce que je voulais faire. (TV)
'When I wanted to pass conservatory, I took advantage of a lunch to tell him what I wanted to do.'
b. Je voulais savoir si tu avais une place demain vers 16h ? (SMS)
'I wanted to know whether you had a place tomorrow around 16h?'

(2) a. Elle a pu rencontrer un beau garçon. (Fiction)
'She could meet a handsome guy.'
b. Tu étais assis très confortablement, tu pouvais rouler dans la neige. (TV)
'You were sitting very comfortably, you could drive in the snow.'

(3) a. Pour prendre soin de toi, j'ai souvent dû délaisser ta sœur aînée. (Letters)
'In order to take care of you, I often had to neglect your elder sister.'
b. Il devait avoir touché une petite fortune pour un tel contrat. (Film)
'He had to make a small fortune for such a contract.'

(4) a. Il nous a bien fallu nous rendre à cette évidence. (Academic)
'We had to acknowledge the evidence.'
b. J'avais de super jambes. Il fallait que j'en profite. (Drama)
'I had nice legs. I had to benefit from this.'

The examples (1–4), taken from the *Corpus de référence du français contemporain* (CRFC; Siepmann et al. 2017), illustrate some essentials of French modal semantics. All of them lexically encode the speaker's attitude to the assertion, modalizing it. The most common typology is the tripartition between deontic (coding authority), epistemic (coding an estimation) and dynamic (coding capacity) modality, with the term *root modality* sometimes being used to cover deontic and dynamic modality (Nuyts 2016). The four French modals presented here seem to allow more or less clear correspondences if considered in isolation: *vouloir* 'want' and impersonal *falloir* 'be necessary' code deontic modality. The modals *devoir* 'must, have to' and *pouvoir* 'can, be able to', however, are inherently polysemous (Boogaart 2009), since they allow deontic/epistemic or even deontic/dynamic/epistemic interpretations respectively.

The picture becomes more blurred if one adds the aspectual dimensions that are coded by grammatical aspect, i. e. *passé composé* (PC) and *imparfait* (IMP). The cases *vouloir* and *falloir* both still express deontic modality in the past, either in form of volition (1a–b) or of necessity (4a–b). Past-tensed *pouvoir* also keeps its polysemy in (2a): it allows a deontic reading (the woman had the permission

to meet the man), a dynamic one (she had the capacity to meet him), and an epistemic one (she may have met him). However, the proposition is ambiguous as to whether the woman has met the man or not. By contrast, the context in (2b) suggests a dynamic reading. Finally, *devoir* in (3a) expresses obligation in the past, whereas in (3b) it construes epistemic modality. French PMCs are thus located at the interface between tense, aspect and modality (TAM) (Desclés 2003). Consequently, it is quite difficult to disentangle their individual semantic values. As will be shown, an aspectual analysis cannot account for the past-tense behaviour of French modals. One appealing proposal is instead put forward by Hacquard (2006), following Bhatt's (1999) seminal work, who notes that perfective aspect on root modals triggers so-called 'actuality entailment': the action has in fact taken place in the actual world. However, the precise reason for this interaction between aspect and modality has been subject to controversy.

Corpus linguistics may contribute another piece in the puzzle by identifying the different form-meaning-correspondences and their usage patterns. If we know how French PMCs are distributed in actual speech, we can derive characteristic patterns and their underlying generalizations from them. It is therefore worth reanalyzing a TAM-phenomenon from a corpus-driven constructional perspective. Our central assumption is that the two French past tenses can be considered an alternation phenomenon: both of them depict a situation in the past, but they do so with different perspectives. Thus, this semantic difference should be reflected in a different lexico-grammatical patterning. We introduce a new reference corpus of French, the CRFC, and perform *distinctive collexeme analysis* (Gries/Stefanowitsch 2004) in order to identify the preferred verbal complements of each PMC, assuming that complementation has a pivotal role in determining their semantics. Concordancing eventually allows the detection of underlying constructional patterns that can be analyzed in terms of common semantic traits.

The paper is structured as follows: Section 2 outlines theoretical considerations, offering an overview of the aspect-modality interface in French. Furthermore, it briefly reviews the body of corpus-based work on French past tenses and gives a sketch on the (corpus-based) construction grammar (CG) paradigm, showing how CG can offer a fresh view on the semantics of aspect and modality. Section 3 introduces the CRFC and describes the methodology of the corpus study. Section 4 contains the results of the corpus analysis. Section 5 discusses the findings. Finally, Section 6 draws a short conclusion.

2 Theoretical considerations

2.1 Aspect and modal verbs in French

The choice between French PC and IMP — similar systems can be found in other Romance languages — is a matter of grammatical aspect. At least since the works of Garey (1957) and Comrie (1976) it has been considered common knowledge that grammatical aspect in French is limited to past tenses: perfective aspect is coded in the PC (and of course in the *passé simple*), construing a situation globally with its temporal boundaries as in (5). By contrast, the IMP codes imperfective aspect and focuses the internal perspective of a situation unfolding in time. Temporal boundaries are not considered, as shown in (6).

(5) Je suis parti de chez moi vers 7h30. (TV)
'I left home at 7h30.'

(6) Nous parlions de ma santé quand soudain ils m'ont annoncé
la mort de mon père. (Diaries & Blogs)
'We were talking about my health when suddenly they announced
the death of my father.'

The PC has relatively clear-cut semantics, encoding the result of an action and depicting this result as one whole event (Desclés/Guentcheva 2003), be it in connection with speech time (resultative) or isolated from speech time (perfective past). That is not the case with the IMP, which can be considered some sort of "chameleon". It can represent aspectual, but also pragmatic and modal meanings: common subsenses are habituality, politeness and counterfactuality. Taking these — and other — usages into account, it has to be asked whether the IMP is in fact a tense, a mood, or a combination of both, whose temporal and modal values are intertwined (cf. Labeau 2002 for an overview of IMP meanings).

Searching for an invariant meaning in monosemic approaches leads to various proposals. Coseriu (1976) sees the core trait of the IMP in its "nonactuality", refuting the claim to assign to it the status of a past tense. According to him, Romance tenses can be broken down to the opposition "actual vs. nonactual": the present tense constitutes the core of the actual level (i.e. an action takes place either in the past, in the present, or in the future), whereas the imperfect is its counterpart, constituting the core of a second, nonactual level (i. e. the realization of the action is somehow impeded and can therefore only be hypothetical). Brisard's (2010) criterion of "virtuality" is similar to this: the IMP creates a second virtual viewpoint distant from the speaker's one. Turning to frameworks that go beyond the sentence level, Weinrich's (1982) concept of discourse grounding is

certainly the most widespread explanation for past-tense use in discourse: the perfective past foregrounds situations, making the plot advance. By contrast, the imperfective backgrounds them, creating periods of stasis (cf. also Michaelis 2011; for a detailed discussion of different approaches to the IMP and related problems cf. Brisard 2010: 487–497).

As evidenced in Section 1, the modal verbs somewhat seem to escape these traditional approaches. Reconsider the examples (1–4): a purely aspectual analysis fails to motivate the past-tense alternations. In (3a), the signal word *souvent* would trigger a habitual reading and thus the IMP. By analogy, the IMP in (1b) is unexpected because the speaker's volition is delimited to the context of conversation. Instead, the polite imperfect in this case encodes a pragmatic value. The past-tense alternation of *vouloir* in (1a) poses another problem because it is not clear why the speaker's volition should be temporally delimited in *j'ai voulu passer le conservatoire*, but undelimited in *ce que je voulais faire*. It does not motivate the one in (4a–b) either. Narrative explanations like discourse grounding also seem to be problematic, since a text linguistic approach cannot be simply adopted to (informal) conversation like in (1a), (2b) or (3b). Briefly put, there must be more to past-tensed modals than the temporal delimitation or the discourse grounding of the proposition.

The works of Bhatt (1999) and Hacquard (2006, 2009) are most notably known for the hypothesis that perfective morphology on root modals neutralizes the modal value of the proposition, replacing it with an uncancelable inference: the proposition takes place in the actual world, giving rise to the so-called 'actuality entailment' effect (cf. Hacquard *to appear* for a detailed comparison of different explanations). Consider the following examples (Hacquard 2009: 288–290):

(7) a. Jane a pu soulever cette table, #mais elle ne l'a pas soulevée.
b. Jane pouvait soulever cette table, mais elle ne l'a pas soulevée.
'Jane was able to lift this table, but she didn't lift it.'

(8) a. Lydia a pu aller chez sa tante (selon les ordres de son père), #mais n'y est pas allée.
b. Lydia pouvait aller chez sa tante (selon les ordres de son père), mais n'y est pas allée.
'Lydia could go to her aunt (according to her father's orders), but she didn't go.'

(9) Bingley a (bien) pu avoir aimé Jane, comme il a (bien) pu ne pas l'aimer.
'Bingley may (well) have loved Jane, just as he may (well) not have loved her.'

It is impossible to cancel the action if perfective aspect operates on a root modal (7a, 8a). This is not the case for epistemic modals as in (9). Imperfective modals as in (7b) and (8b) are not subject to actuality entailment, which is presumably due to their generic nature. Note that whereas the modals *pouvoir*, *devoir* and *vouloir* have been analyzed fairly extensively in this framework, to our knowledge impersonal *falloir* has not been taken into consideration yet. It may be assumed that actuality entailment also affects perfective *falloir*, as it equally codes deontic modality. The following section deals with how corpus-driven CG can complement these theoretical claims with empirical data and how the semantics of PMCs can be grasped in terms of lexico-grammatical constructions.

2.2 Corpora, constructions and usage

The body of corpus-based work on French past tenses is rather modest, since several reasons reduce the representativeness of the studies. First of all, the data is based on small corpora sometimes representing particular text types. Common genres investigated are newspapers (Waugh/Monville-Burston 1986), television talk (Labeau 2006), sports commentaries (Labeau 2004, 2007) or obituaries (Do-Hurinville 2010, Labeau 2013). While this is not a lacuna *per se*, it would certainly be appreciated if the data basis were to be expanded to bigger sample sizes of spoken informal varieties. Unsurprisingly, the call for a mega-corpus of contemporary French has been repeatedly issued (Deulofeu/Debaisieux 2012, Bilger/Cappeau 2013). Second, these studies provide descriptive frequencies or percentages, without using any sort of inferential statistics that could possibly generalize the findings. Third, they mostly do not give any insights into whether single verbs show preferences for one of the past tenses, which could shed further light on the relationship between lexical and grammatical aspect. Narrowing the focus down to modal verbs, the only study providing frequency data is Blumenthal (1976): French PMCs prefer to be realized imperfectively, be it in radio interviews (ratio IMP to PC 2:1) or in fiction (ratio 3.5:1), the exception being newspaper articles (ratio approx. 1:1). But similar to the aforementioned studies, the sample size (no past-tense construction occurs more than 100 times) does not permit any representativeness.

Large-scale corpus linguistic work has shown that language consists of more or less schematized form-meaning-correspondences, so-called *constructions*: a linguistic unit is stored as a construction as long as it has non-compositional semantics (Goldberg 1995) or as long as it occurs with sufficient frequency (Goldberg 2006). Most constructional theories are also usage-based, highlighting the importance of frequency in language structure and acquisition: a given category is made of some high-frequent prototypes and a large number of low-frequent peripheral members (Diessel 2015). Prototypes are processed faster and

can trigger priming effects, facilitating the acquisition of peripheral members (Ellis 2002). Another major CG tenet concerns the inseparability of lexis and grammar (Römer 2009, Hunston 2015), be it from a *lexis-to-grammar* perspective (a linguistic unit selects lexico-grammatical environments in which it occurs preferably), or from a *grammar-to-lexis* perspective (a grammatical construction attracts specific collocates). One statistically reliable method to calculate the attraction between a construction and its collocations is collostructional analysis (Gries/Stefanowitsch 2004). Taking observed and expected frequencies of collocates into account, one can calculate the collostructional strength, a value indicating how strongly a construction attracts a collocate in a slot. These so-called *collexemes* can be ranked in terms of their collostructional strength, with highly distinctive collexemes being indicative of relatively frozen constructional patterns, prone to be entrenched and stored separately.

What do these ideas imply for the analysis of French PMCs? We assume that they can be analyzed from a corpus-driven constructional perspective. Perfective and imperfective modal constructions are an alternation phenomenon and thus likely to co-occur with different sets of verbal complements. The retrieval of distinctive collexemes may shed light on preferred co-occurrence patterns and eventually on the underlying semantics of PMCs. Notions such as 'actuality entailment' can thus possibly be grasped in terms of highly frequent lexico-grammatical constructions. In fact, there is substantial empirical evidence that a CG analysis of modality is possible and explanatory (see e. g. the thematic issue 8/1 of *Constructions and Frames*). Consider, for instance, syntactic patterns in English that correlate with epistemic modality (Wärnsby 2002) or collocational preferences of modal verbs (Hilpert 2016). Further hints at the constructional relevance of modal semantics are provided by De Haan (2012), who investigates the patterning of the English modal *must*. His findings show a strong correlation between modality and verbal construction: '*must* + progressive' as well as '*must* + perfect' almost exclusively express epistemic modality, whereas '*must* + V' yields a deontic interpretation. Moreover, he points out the importance of register and person as additional factors. Another piece of evidence comes from German and Dutch, where impersonal complementation triggers an epistemic reading (Boogaart/Fortuin 2016: 529f.).

3 Corpus and methodology

3.1 The Corpus de référence du français contemporain

The CRFC is the first genre-diverse reference corpus of contemporary French with about 310 million words, evenly distributed among spoken, written and

pseudo-spoken[1] varieties (cf. Siepmann et al. 2017 for detailed information on its design and compilation). The corpus has been POS-tagged via the *French Tree Tagger* (Stein 2003), but lacks prosodic annotation for the spoken varieties as well as syntactic parsing. It includes over 155 million words of (pseudo-)spoken language such as informal conversation, drama scripts, discussion forums, chats or television subtitles. The written subcorpora include another 155 million words of academic texts and lectures, prose fiction, newspaper articles, parliament speeches and several smaller-sized genres such as diaries and blogs. Table 1 illustrates its composition.

Table 1: Compilation of the CRFC (Siepmann et al. 2017: 70).

Category	Subcorpus	Size in mill.	Category	Subcorpus	Size in mill.
Spoken	Informal	30	Written	Academic papers	30
Pseudo-spoken	Drama scripts	30		Non-academic texts	30
	TV subtitles	2,5		Prose fiction	30
	SMS and chats	2,5		Newspaper articles	45
	Discussion forums	60		Magazines	10
Pseudo-written	Formal	30		Diaries and blogs	5
				Letters and e-mails	1
				Miscellaneous	4
		155			155

Previous studies in lexicography (Siepmann 2015) and descriptive grammar (Siepmann/Bürgel 2015, 2016) have shown that a thorough corpus-driven look at linguistic phenomena in French can generate new insights on their distributions that have hitherto been neglected in traditional grammars. The CRFC is currently available on-demand on the platform *Sketch Engine*.

3.2 Data retrieval

The PMC can roughly be schematized as a tripartite structure [SUBJ MOD_{PST} COMP], with the subject being followed by the past-tensed modal and the verbal complementation slot, see (10). Note that two optional slots are added in order to account for possible adverbs or clitics. The examples (11a–d) show different instantiated constructs.

1 The term *pseudo-spoken* can best be explained in terms of the well-known distinction between *immediacy* and *distance* elaborated by Koch/Oesterreicher (²2011), referring to written language that typically exhibits spoken language characteristics, e. g. chats, text messages and threads in discussion forums (*immediate language*).

(10) [SUBJ (OPT) MOD_{PST} (OPT) COMP]

(11)	a.	Je		voulais	te	demander
	b.	Mon frère	n'a pas	pu		rentrer
	c.	Il	lui	fallait	bientôt	arriver
	d.	On	ne	devait	plus jamais	retourner

1. **Filtering and CQL-commands**: With the corresponding CQL-commands, it was possible to obtain all instances of PMCs. The verbal complements were retrieved in the interval 0–3R. As for the IMP-constructions, the conjunction *si* in the left periphery (0–5L) was filtered out in order to avoid irrealis conditional clauses, where the conjunction would automatically trigger the IMP (*Si tu faisais* 'If you did'). One remaining problem concerned the occurrence of *que* 'that' in the left context, as it triggers the *imparfait de concordance* if the matrix verb is realized in a past tense, e. g. *Il a dit que je pouvais venir* 'He said that I could come'. This bias could not be eliminated.
2. **Collecting raw frequencies:** In a first step, we listed the most frequent verbal complements of each PMC, resulting in lists of 50 verbs for each PMC.[2] As Gries et al. (2010) have pointed out, however, raw frequencies are not reliable enough to tell whether a complement has a preference for one of the two constructions. That is why a distinctive collexeme analysis was performed.
3. The **distinctive collexeme analysis** was carried out with the R script *Coll. Analysis* 3.2a (Gries 2007). Following previous work, the lists only present the most distinctive collexemes, in our case 15 (it will be noted if there are any more distinctive collexemes). One reason for this can be formulated from a statistical point of view: the significance level was put at $p < 0.001$, corresponding to a collostructional strength of over 3. Keep in mind that the complementation slot is open and every verb in the lexicon could theoretically occur in it. Consequently, a certain number of verbal complements would gain at least significant collostructional strength (Coll.str. > 1.3; $p < 0.05$), blurring the overall picture. By analogy, focussing on the most distinctive collexemes permits a) the detection of frozen phraseological expressions and b) a better evaluation of possible links between modality and constructional patterning: in the usage-based CG framework, high-frequent items are also often prototypes. If these prototypes, in our case the most distinctive collexemes, occur in a certain PMC then it is reasonable to assume that they form a close semantic

2 One might object that low-frequent verbs could possibly have high collostructional strength values. This was not the case, as has been tested for several examples: no significant collostructional strength could be measured.

link with the construction, following the basic principle of corpus linguistics that items occurring in similar contexts also have similar semantics (Stubbs 2016).

4. **Concordancing:** The fourth step was to investigate the constructional patterns of the PMCs by means of concordancing. The verbal complement may be, for instance, embedded in a secondary pattern; correlations may be established between modality and constructional patterning. This step, however, was only performed for the most distinctive collexemes and involved the researcher's intuition and a more qualitative analysis. Of course, future studies can apply more refined methods, submitting manually coded instances to multivariate procedures like correspondence analysis to identify semantic clusters; but in this case, and due to space restrictions, the results should still provide sufficiently clear answers to our questions.

4. Results

4.1 Overview: Raw frequencies

In a first step, we give an overview of frequencies and distributions. Figure 1 shows the raw frequencies for the four French PMCs in the CRFC.

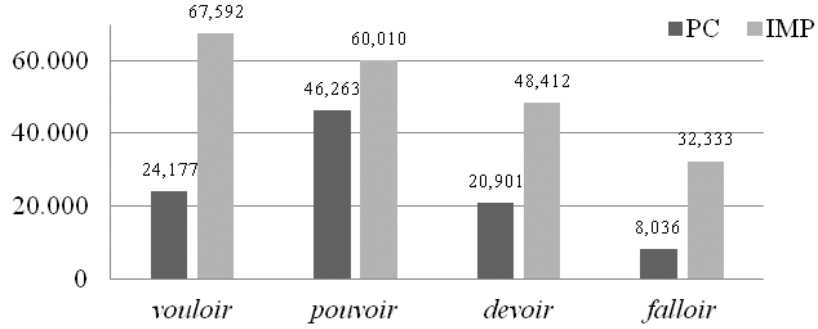


Figure 1: Raw frequencies of the French PMCs in the CRFC.

First of all, it has to be noted that raw frequencies of up to more than 60,000 are reached. Unsurprisingly, all of the PMCs tend to be realized imperfectively. The overall picture, however, is not uniform. The verb *pouvoir* almost equalizes this difference, with the IMP being only 1.3 times as frequent as the PC. The differences become bigger with *devoir* and *vouloir*, rising up to four times as frequent

imperfects with *falloir*. Furthermore, the ratio between IMP and PC is subject to genre-specific differences, as shown in Table 2. The IMP is only half as frequent as the PC in letters and e-mails, although this finding has to be treated with caution due to the small corpus size. In almost all the other subcorpora the IMP is more frequent, reaching the biggest ratio differences in the SMS and prose fiction subcorpora.

Table 2: Genre-specific IMP to PC ratios.

Subcorpus	IMP : PC	Subcorpus	IMP : PC
Letters and e-Mails	0.46	TV	1.89
Miscellaneous	0.66	Non-academic	1.94
Newspapers	1.17	Drama	2.19
Spoken formal	1.33	Discussion forums	2.24
Magazines	1.66	Spoken informal	2.38
Diaries and blogs	1.67	SMS	4.81
Academic	1.79	Prose fiction	5.77

Note finally that the genre differences already mentioned by Blumenthal (1976) are mirrored in the CRFC. On the one hand, we can observe a balance between the two forms in newspapers (ratio 1.17), on the other hand, the biggest difference can also be found in prose fiction (ratio 5.77).

4.2 Collostructional analysis: distinctive collexemes and their usage patterns

This section now turns to the results of the distinctive collexeme analysis. Of primary concern is the question of how the distributional properties of French PMCs can be described and if it is possible to group the collexemes into semantic classes and lexico-grammatical patterns.

The first case study is dedicated to the verb *pouvoir*. Table 3 compares the most distinctive collexemes for the two constructions. The PC seems to prefer verbs that represent some sort of (visual) realization such as *constater* 'state, notice', *découvrir* 'discover', *voir* 'see' or *observer* 'observe'. Additionally, all of the collexemes are transitive verbs. The most distinctive collexeme *constater* has infinite collostructional strength, which hints at a frozen pattern (the concrete value could not be calculated due to processor restrictions). Quite strikingly, not a single distinctive PC-collexeme is represented in the IMP-list. Verbs such as *espérer* 'hope', *imaginer* 'imagine', *savoir* 'know', *supporter* 'bear' or *penser* 'think' could possibly be grouped into a class of cognitive/psych-verbs.

Table 3: Distinctive collexemes of the past-tense constructions of *pouvoir*.

PC (N = 46,263)		IMP (N = 60,010)	
Collexeme	Coll.Str.	Collexeme	Coll.Str.
constater 'state, notice'	Inf	*espérer* 'hope'	81.77
découvrir 'discover'	80.74	*imaginer* 'imagine'	61.54
voir 'see'	59.92	*savoir* 'know'	60.84
observer 'observe'	55.41	*durer* 'last'	46.47
mesurer 'measure'	54.35	*permettre* 'allow'	37.55
apprécier 'appreciate'	53.86	*avoir* 'have'	26.84
montrer 'show'	53.44	*laisser* 'let'	26.19
lire 'read'	50.11	*être* 'be'	22.08
assister 'assist'	45.34	*continuer* 'continue'	10.27
développer 'develop'	40.57	*supporter* 'bear'	9.83
obtenir 'obtain'	38.17	*penser* 'think'	9.79
établir 'establish'	37.47	*aller* 'go'	7.46
rencontrer 'meet'	36.47	*compter* 'count'	7.05
résister 'resist'	29.08	*arriver* 'arrive'	6.58
profiter 'profit'	25.67	*manger* 'eat'	5.73
14 others			

Concordancing reveals clear-cut patterns for the PMCs of *pouvoir* (see Figure 2). The PC-constructions describe how the subject has managed to realize something. Actuality entailment is at hand, as the instances can be substituted with their lesser marked equivalent without a modal, e. g. *j'ai constaté que*. Two aspects hint at the phraseological nature of the PC-construction: first, they have a much higher number of distinctive collexemes than the IMP (29 versus 15); second, they occur predominantly with 1SG and 1PL. The IMP, by contrast, seems to express dynamic modality with constructs such as *on pouvait imaginer que* 'one could imagine that' or *je (ne) pouvais savoir que* 'I could (not) know that'.

The analysis of *vouloir* reveals another case of differently distributed collexemes (see Table 4). The IMP-collexemes can be grouped into a class of 'discourse verbs' with the members *dire* 'say', *savoir* 'know', *demander* 'ask', *parler* 'talk' and *remercier* 'thank'. The concordance in Figure 3 shows that they are used with the polite imperfect. The IMP almost exclusively instantiates this subsense as the overall number of distinctive collexemes (12) is rather low.

The PC differs insofar as none of its collexemes has an extraordinarily high collostructional strength, the most distinctive ones being *prendre* 'take' and *faire* 'make, do'. This in turn would mean that the PC of *vouloir* allows freer combinations. In fact, it simply seems to express the literal meaning of volition, as shown in Figure 3. The contrast between the PMCs of *vouloir* can also be described by means of 'speech situation': the IMP-construction instantiates a polite use and is thus predominant in dialogical settings, e. g. if the speaker addresses his

Drama	de Duchemin. RICHARD: Comme vous	avez pu	le *constater* nous venons de pratiquer
Fiction	sont de redoutables chicaneurs, j’	ai pu	le *constater* quand j’ai eu affaire à
Formal	surfaces sont équilibrés ? Nous	avons pu	le *constater* récemment avec les
TV	– Magique, bien sûr, mais je n’	ai pas pu	le *voir* naître car j’ai eu une
Forums	les sentiers et chemins parcourus j’	ai pu	*mesurer* avec effroi l’ampleur de la
Formal	Bas, en Suède et en Espagne. Nous	avons pu	*voir*, je le souligne, toute la
Forums	dire plus sur ce reportage que je n’	ai pas pu	*voir*, mais pour ne pas encombrer le
Formal	schéma interrégional du littoral n’	a jamais pu	*voir* le jour et qu’aucun schéma de
TV	Car par ma propre pratique, j’	ai pu	*voir* les qualités de Bokar Rimpoché
TV	l’influence qu’on souhaiterait et qu’on	a pu	*voir* par le passé à d’autres
TV	à qui raconterait la meilleure. Ca	pouvait	*durer* 45 mn. Ils avaient tous les 2 le
Fiction	pessimiste sur le résultat que l’on	pouvait	en *espérer*. Il me l’avait dit
Magaz	que c’était le meilleur accord qu’on	pouvait	*espérer*. “Elle fut en effet à deux
News	mètres. “Ca dépasse tout ce que l’on	pouvait	*imaginer*, vu d’ici”, soupirent Sophy
News	par la mère. Sauf que cela ne	pouvait pas	*durer* éternellement”, commente-en
Fiction	c’était une fatalité et qu’elle ne	pouvait pas	*espérer* mieux. - Écrase-le, dit-elle
TV	les cheveux, il est trop moche! – On	pouvait pas	*savoir* ! Il est si moche que ça? Paola
Drama	toi. Oui... mais je savais pas, je	pouvais pas	*savoir*... J’étais à un feu... Non.
Drama	sur le champ. Le militaire: Vous ne	pouviez pas	*savoir* à qui vous aviez affaire.
Fiction	, que ça lui était égal. Elle ne	pouvait pas	*savoir* que c’était mon emploi qui

Figure 2: Concordances of the PMCs of *pouvoir.*

Table 4: Distinctive collexemes of the past-tense constructions of *vouloir*.

PC (N = 24,177)		IMP (N = 67,592)	
Collexeme	Coll.Str.	Collexeme	Coll.Str.
prendre 'take'	25.52	*dire* 'say'	174.38
faire 'make, do'	25.51	*savoir* 'know'	108.69
mettre 'put'	24.44	*demander* 'ask'	44.72
donner 'give'	24.40	*parler* 'talk'	31.32
prêter 'lend'	18.02	*être* 'be'	21.01
créer 'create'	14.92	*remercier* 'thank'	15.65
reprendre 'regain, start again'	14.68	*avoir* 'have'	7.53
croire 'believe'	14.29	*vivre* 'live'	5.40
montrer 'show'	13.92	*rester* 'stay'	5.23
répondre 'answer'	13.52	*devenir* 'become'	4.92
comprendre 'understand'	8.68	*voir* 'see'	3.86
rendre 'give back'	8.08	*entendre* 'hear'	3.20
essayer 'try'	7.18		
jouer 'play'	6.67		
tuer 'kill'	6.16		
2 others			

interlocutor(s). The PC-construction in turn is mainly used in narrative settings. These perfective instances do not code whether the action has taken place or not; actuality entailment is not triggered (cf. Hacquard *to appear* for a formal explanation).

The analysis of *falloir* reveals two considerably different collexeme sets (see Table 5). The list of PC-collexemes is characterized by a steep falling curve: the most distinctive complement *attendre* 'wait' has an extraordinarily high collostructional strength, followed in second rank by *adapter* 'adapt', whose value is more than ten times lower. Furthermore, most of them are rather middle-/low-frequent verbs, e. g. *réapprendre* 'relearn' or *batailler* 'fight'. Interestingly, some of the collexemes can be grouped into a class of 'construction'-verbs, such as *inventer* 'invent', *refaire* 'redo', *reconstruire* 'reconstruct', *créer* 'create' and *composer* 'compose'.

Most IMP-collexemes, on the other hand, are high-frequent verbs that cannot be easily grouped into a coherent semantic class. However, the concordance in Figure 4 gives evidence of an entrenched pattern *il a fallu attendre* + 'event / date' related to a narrative-historical text type. This pattern, as well as all the other examples, infers actuality entailment. On the contrary, typical IMP-constructs such as (*il*) *fallait y penser* / *le dire* 'should have thought about it / have said it' share a counterfactual meaning. The construct (*il*) *fallait le faire* 'it was necessary to do it', however, may also be non-implicative, as some contexts do not necessarily involve the fulfilment of the action.

Formal	Non pour la détruire, comme certains	ont voulu	le *faire* croire, mais pour garder
TV	sort. Et de me demander pourquoi, j’	ai voulu	lui *faire* honneur, faire le mieux
Drama	plus jeune âge... Quand ma belle mère	a voulu	me *faire* assassiner car j’étais
Forums	d’économie! Du moins c’est ce qu’elle	a voulu	me *faire* croire il y’a quelques
TV	Pour faire rire, c’est plutôt Moi j’	ai voulu	*mettre* la main sur Robin ne
Drama	pensé que c’était un souvenir, ils	n’ont pas voulu	nous *faire* de peine. LUDOVIC:
TV	au Président et à son – P. Loison: Il	a voulu	*prendre* le pouls du pays,
Forums	es en obesité morbide alors? Oui j’	ai voulu	que *prendre* vite les rdv pour m’en
Forums	terrestre, ni mortel ni immortel, j’	ai voulu	te *donner* le pouvoir de te former
TV	même. Ils seraient ravis mais n’	ont pas voulu	vous *faire* de la peine. Pour
Forums	coté de ça je suis pas fatiguée.. dc je	voulais	*savoir* si ça pouvait qd meme venir
Forums	acheter de la proteine en poudre et je	voulais	*savoir* laquelle choisir entre
Forums	merci de vos conseils!!!! Etrebien2 je	voulais	*savoir* si a 48kg pou 1.61 si tes
Forums	un peu de beurre mais c’est rare Et je	voulais	*savoir* si je pouvais enlever le
Forums	la boite il y a marquer 4-6 mois mes je	voulais	*savoir* si des maman avez déjà
Forums	manger (ce qui en fait es faux). Je	voulais	*dire* que, si par exemple, la matin
Forums	: 02/07/02 à 18h46 Contenu: Salut, Je	voulais	te *dire* que les trente secondes
Forums	au Nord, l’autre Sud. Bien réussi, tu	voulais	*dire* ---------- 21/6/2001 –
Drama	ai pas fait déranger. Dites donc, je	voulais	vous *demander*, vous en avez pas
Drama	depuis tout à l’heure ??? Antonia: Je	voulais	*savoir* si vous en vouliez une ou

Figure 3: Concordances of the PMCs of *vouloir.*

Table 5: Distinctive collexemes of the past-tense constructions of *falloir*.

PC (N = 8,046)		IMP (N = 32,333)	
Collexeme	Coll Str.	Collexeme	Coll.Str.
attendre 'wait'	195.42	*faire* 'make, do'	78.73
adapter 'adapt'	13.95	*dire* 'say'	37.58
convaincre 'convince'	10.01	*voir* 'see'	32.23
battre 'fight'	9.81	*être* 'be'	23.02
inventer 'invent'	8.64	*penser* 'think'	21.87
réapprendre 'relearn'	7.36	*laisser* 'let'	15.50
refaire 'redo, remake'	6.47	*oser* 'dare'	12.84
batailler 'fight'	5.40	*prendre* 'take'	10.83
reconstruire 'reconstruct'	5.32	*lire* 'read'	9.38
créer 'create'	5.14	*parler* 'talk'	7.61
apprendre 'learn'	4.84	*éviter* 'avoid'	7.24
gérer 'manage, handle'	4.56	*aller* 'go'	7.17
expliquer 'explain'	4.51	*donner* 'give'	7.03
composer 'compose'	4.21	*compter* 'count'	6.46
résoudre 'solve'	3.88	*arrêter* 'stop'	5.35
4 others		2 others	

The analysis of the PMCs of *devoir* is presented last, since it demonstrates the limits of a collostructional analysis limited to complementation. Several observations can be made for the PC (see Table 6). First, almost all collexemes are telic. Second, they tend to be middle-/low-frequent, e. g. verbs such as *abandonner* 'abandon', *résoudre* 'solve' or *affronter* 'face'. Third, collexemes such as *tromper* 'be mistaken, wrong (refl.); cheat (tr.)', *renoncer* 'give up', *abandonner* 'abandon' or *subir* 'suffer' hint at negative semantic prosody (Louw 1993). It is, however, difficult to tell which type of modality is being expressed by the PC. The same difficulties hold for the IMP-constructions. The strongest collexeme *être* can be explained with its use as the passive auxiliary, indicating that the imperfective constructions of *devoir* tend to be realized with the passive voice.

A closer look at concordances can help refine the picture (see Figure 5). The PC-constructions can be clustered into two classes: first, the constructions that express epistemic modality with constructs such as *j'ai dû me tromper* 'I must have been wrong' or its less frequent variant *j'ai dû oublier* 'I must have forgotten'. Second, the constructions that expresses deontic modality, where the subject is forced to react to external circumstances. Prominent patterns are *il a dû renoncer* 'he had to give up', *elle a dû s'adapter* 'she had to adapt' or *on a dû quitter* 'we had to leave', all of them triggering actuality entailment. By contrast, the IMP seems to have a preference for epistemic use, especially in combination with *avoir* or *être*.

News	l’enchaînement des faillites. Il	a fallu	*attendre* fin 2011 pour que l’idée d
News	French; Français Graphique: Il	a fallu	*attendre* 22 ans avant de pouvoir
Non-ac	Widerberg semble tarder Il	a fallu	*attendre* les années 1970 et la
Non-ac	fait de leur teneur en phosphore. Il	a fallu	*attendre* l’application du procédé
Non-ac	est fondé à se demander pourquoi il	a fallu	*attendre* trois ans, et une campagne
Non-ac	“La germanicité” Il	a fallu	*attendre* Madame de Staël et son
Non-ac	phraste (370-285 av. J.-C.) mais il	a fallu	*attendre* Adolphe Brongniard (1829)
Formal	mage, oui ! M. Jean-Luc Drapeau. Il	a donc fallu	*attendre* que la gauche arrive au
Formal	ces derniers jours à la Réunion. Il	a fallu	*attendre* sept jours et la perte de
Formal	déloyaux. Malheureusement, il n’	a pas fallu	*attendre* longtemps pour voir le
TV	, c’était pas le tout de sortir, il	fallait	*faire* son trou. MITTERAND! MITTERA
Inform	écrivait! Mais cette communauté il	fallait	la *faire*, ce qui exigeait qu’on
Drama	: (l’ignore) L’oncle Philippe. Mais	fallait	le *dire*, Steevy m’a parlé au moins
Film	C’est une americaine qui le tient. –	Fallait	le *dire*, ouais ! 197 00: 17:06
Drama	sur les lacets de son bustier) Il	fallait	le *dire* avant. Les toilettes t sont
Forums	soins que je lui prodiguat etc... il	fallait	me *faire* lacher mon combat de mère!
News	Jean-Jacques” , tout simplement, il	fallait	*oser*. D’autant plus en cette
Forums	, ce sont nos emprunts qu’il ne	fallait pas	*faire*. Vous ne pouvez pas vivre
Fiction	s’en apercevait. Sauf eux. Mais il	fallait	*faire* attention à ne pas se
TV	, voilà. – Ca vous a appris qu’il	fallait	*faire* gaffe au thème choisi? – Oui,

Figure 4: Concordances of the PMCs of *falloir.*

Table 6: Distinctive collexemes of the past-tense constructions of *devoir*.

PC (N = 20,901)		IMP (N = 48,412)	
Collexeme	Coll.Str.	Collexeme	Coll.Str.
tromper 'be wrong; cheat'	73.03	*être* 'be'	250.35
renoncer 'give up, renounce'	60.36	*avoir* 'have'	91.26
abandonner 'abandon'	46.07	*permettre* 'allow'	61.93
faire 'make, do'	45.38	*conduire* 'lead'	26.64
adapter 'adapt'	38.75	*rester* 'stay'	23.78
oublier 'forget'	37.94	*savoir* 'know'	23.77
quitter 'leave'	33.21	*devenir* 'become'	19.41
arrêter 'stop'	31.56	*servir* 'serve'	18.48
entendre 'hear'	26.95	*arriver* 'arrive'	16.17
résoudre 'solve'	26.76	*durer* 'last'	14.93
subir 'suffer'	25.85	*retrouver* 'find; meet'	13.80
fermer 'close'	24.92	*tenir* 'hold'	11.10
affronter 'face'	24.48	*donner* 'give'	10.17
dire 'say'	22.09	*revenir* 'return'	6.64
tomber 'fall'	20.13	*aller* 'go'	6.39
2 others		3 others	

Some cases found in the corpus, however, may allow a deontic interpretation as in (12) and (13). Yet, they do not imply that the action has in fact taken place.

(12) On pourrait dire que, pour nous, la recherche universitaire devait absolument être liée au mouvement social. (Discussion Forums)
'You could say that for us, scientific research had to be absolutely related to the social movement.'

(13) Celui-ci [le dossier, OW] devait être remis à la Commission européenne mercredi. (Newspapers)
'The dossier had to be handed to the European Commission on Wednesday.'

5 Modal constructions?

The findings of the corpus-driven analysis of French PMCs can be summarized in Table 7. The empirical evidence suggests that it is possible to establish certain correlations between modality and constructional patterning: the PC of *falloir* and *pouvoir* infers actuality entailment; perfective *vouloir* expresses literal volition, and *devoir* remains ambiguous between epistemic interpretations and actuality entailment with deontic modality. The IMP, on the other hand, encodes

TV	temps, mais dans les années 60, elle	a dû	*faire* face à une sacrée concurrence: les
Drama	francs de bougies! Bûchette: Ca	a dû	*faire* plaisir au vieux barbu... C'était
TV	ce travail à plein temps. Là, j'	ai dû	*faire* un choix, peut-être le plus grand
Drama	après le top... Top!" PAUL – J'	ai dû	*faire* une erreur. Au revoir. GASTON - Au
Drama	parles de mémé. LUDOVIC – Oui, elle	a dû	leur *faire* peur. CLEMENCE – Mais non
Film	fric... Je vais voir en cuisine. J'	ai dû	me *tromper* d'adresse. Des canalisations
Drama	non, c'est Jonathan ? Excusez-moi, j'	ai dû	me tromper de numéro... Non, je me suis
TV	de 24 ans sa cadette, Moustapha	a dû	se *faire* à l'idée de côtoyer des gens
Fiction	il a sauté du balcon, l'atterissage	a dû	se *faire* en douceur car il n'a été suivi
TV	incohérences. Retournez l'interroger. Il	a dû	se *tromper*. C'est déjà un miracle
Fiction	, un soir, devant un caboulot – je	devais	*avoir* cinq ou six ans – qui rampait sur
Drama	venue ici que pour le tableau? Il	devait	*avoir* de la valeur... a présent je
Fiction	. Je m'aperçus que cette chose, qui	devait	*avoir* soixante-dix ans, ne portait pas
Fiction	présence du blessé mais le Stéphane	devait	*avoir* une sorte de sixième sens car comme
Drama	le saviez. Nathalie – Non, et je	devais	bien être la seule, vous avez bien du vous
Forums	etc.. Si tout ce qui est légitime	devait	être accepté ss évolution, on en serait
TV	cette chaîne. La bombe cachée dedans	devait	être de fabrication artisanale. – C'est
TV	informatiques, et sinon, cela	devait	être détruit sans laisser de traces. C'
Fiction	et les éclopés étaient demeurés. Ils	devaient	être évacués sur Bangkok par le prochain
Inform	spécifique comme ça <Who nb="1"/> ça	devait	être la journée <Who nb="2"/> hm hm une

Figure 5: Concordances of the PMCs of *devoir.*

counterfactual or non-implicative actions as with *falloir* or politeness with *vouloir*. Imperfective *pouvoir* predominantly expresses dynamic modality, whereas *devoir* is ambiguous: it is mainly used for assumptions in the past, but permits deontic readings as well.

Table 7: Overview of the French PMCs, their semantics and prototypical patterns.

	PC	**IMP**
falloir	deontic: actuality entailment (*il a fallu attendre* + 'event')	counterfactual (*[il] fallait oser / y penser*) non-implicative (*il fallait dire*)
devoir	epistemic (*j'ai dû me tromper*) actuality entailment (*il a dû quitter*)	epistemic (*cela devait permettre / il devait avoir*) deontic (*qqc devait être + past participle*)
pouvoir	dynamic: actuality entailment (*j'ai pu*+ 'realization')	dynamic (*je (ne) pouvais* + 'imaginer')
vouloir	deontic: volition (*j'ai voulu prendre*)	polite imperfect (*je voulais dire que*)

The major question now is whether we can call the structures investigated genuine *constructions* at all. Recall that this term was *a priori* used as a tool to grasp the tripartite string of subject, past-tensed modal and verbal complement in the corpus analysis. Yet, it appears reasonable to treat them as constructions in a narrower sense. They fulfil the criterion of non-compositionality because it is often only the context of the assertion that disambiguates the modality coded in it. This is especially true for all those PMCs that trigger actuality entailment. The effect is simply not predictable from the mere combination of perfective aspect and root modal. Take, for instance, the string *il a dû renoncer* that can be read with an epistemic meaning (he must have given up) or with actuality entailment (he had to give up). Likewise, it is, strictly speaking, impossible to deduce the polite use of imperfective *vouloir* just from the linear sequence of the string *je voulais dire que*. The second criterion, sufficient frequency, is also met. This has been shown by the usage-based collostructional analysis: French PMCs do not select verbal complements arbitrarily but attract specific collexemes with very high frequencies. The most distinctive collexemes often appear in fixed expressions such as *j'ai pu constater que*, *je voulais dire que*, *je voulais savoir si*, *fallait y penser* or *j'ai dû me tromper*.

Finally, some remarks on the methodology adopted in this study: it should have become clear that a collostructional approach can generate interesting findings even if the alternation between French past tenses cannot strictly be seen as synonymous. Yet, it confirms the assumption that different semantics imply different lexico-grammatical patterns. Moreover, the corpus-driven approach permits us to disentangle the individual semantic values encoded in a construction

by extracting the most frequent instantiations, complementing the claims made by theoretical linguists. Only the modal *devoir* remains an outstanding problem, which should encourage further empirical investigations. Note, eventually, that the collostructional analysis has been restricted to complementation patterns. Analyzing modality, of course, necessitates the consideration of additional factors such as the use of pronouns or the role of negation, which is out of scope of this study and should be addressed in follow-up studies.

6 Conclusion

The present contribution aimed at demonstrating how a corpus-driven approach to French PMCs can reveal new insights into their distributional properties and thus into their semantics. The point of departure was to tackle a phenomenon that has received considerable attention in formal linguistics, but whose description in large-scale corpora had been a shortcoming. By means of distinctive collexeme analysis and concordancing, it has been possible to extract respective form-function-correspondences of the four past-tensed modals. The PMCs could thus be described as lexico-grammatical constructions that attract specific collexemes. The co-occurrence patterns reflect underlying semantics and demonstrate that PMCs can encode a variety of meanings that are not restricted to modality, providing an argument for a CG approach. Finally, the theoretical claims related to the actuality entailment effect could be validated on empirical grounds. It is hoped that our investigation stimulates further research into the link between lexico-grammatical patterns and the underlying semantics of French PMCs.

References

Bhatt, Rajesh. 1999. *Covert modality in non-finite contexts*. Ph.D. Thesis, Univ. of Pennsylvania.

Bilger, Mireille and Paul Cappeau. 2013. Comment les données de corpus pourraient renouveler les manuels de grammaire? *Linx* 68/69: 177–199.

Blumenthal, Peter. 1976. Imperfekt und Perfekt der französischen Modalverben. *Zeitschrift für französische Sprache und Literatur* 86/1: 26–39.

Boogaart, Ronny. 2004. Aspect and Aktionsart. In Geert Booij and Herbert Wiegand (eds.), *Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung*, 1165–1180. Berlin: de Gruyter.

Boogaart, Ronny. 2009. Semantics and pragmatics in construction grammar: The case of modal verbs. In Alexander Bergs and Gabriele Diewald (eds.), *Contexts and constructions*, 213–241. Amsterdam: Benjamins.

Boogaart, Ronny and Egbert Fortuin. 2016. Modality and Mood in Cognitive Linguistics and Construction Grammars. In Jan Nuyts and Johan van der Auwera (eds.), *The Oxford Handbook of Modality and Mood*, 514–534. Oxford: OUP.

Borgonovo, Claudia and Sarah Cummins, 2007. Tensed modals. In Luis Eguren and Olga Fernández-Soriano (eds.), *Coreference, Modality, and Focus: Studies on the Syntax-Semantics Interface*, 1–18. Amsterdam: Benjamins.

Brisard, Frank. 2010. Aspects of virtuality in the meaning of the French imparfait. *Linguistics* 48/2: 487–524.

Cappelle, Bert and Ilse Depraetere. 2016. Modal Meaning in Construction Grammar. *Constructions and Frames* 8/1: 1–6.

Comrie, Bernard. 1976. *Aspect*. Cambridge: CUP.

Coseriu, Eugenio. 1976. *Das romanische Verbalsystem*. Tübingen: Narr.

De Haan, Ferdinand. 2012. The Relevance of Constructions for the Interpretation of Modal Meaning: The Case of Must. *English Studies* 93/6: 700–728.

Desclés, Jean-Pierre. 2003. Interactions entre les valeurs de pouvoir, vouloir, devoir. In Merete Birkelund et al. (eds.), *Aspects de la modalité*, 49–67. Tübingen: Max Niemeyer.

Desclés, Jean-Pierre. and Zlatka Guentchéva. 2003. Comment déterminer les significations du passé composé par une exploration contextuelle? *Langue française* 138/1: 48–60.

Deulofeu, Henri-José and Jeanne-Marie Debaisieux. 2012. Une tâche à accomplir pour la linguistique française du XXI^{e} siècle: élaborer une grammaire des usages du français. *Langue française* 176: 27–46.

Diessel, Holger. 2015. Usage-based construction grammar. In Eva Dąbrowska and Dagmar Divjak (eds.), *Handbook of cognitive linguistics*, 296–322. Berlin: De Gruyter Mouton.

Do-Hurinville, Danh Thành. 2010. Etude des temps verbaux dans les articles nécrologiques. *Syntaxe & Sémantique* 11: 83–111.

Ellis, Nick. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24/2: 143–188.

Garey, Howard B. 1957. Verbal aspect in French. *Language* 33/2: 91–110.

Goldberg, Adele. 1995. *Constructions: a construction grammar approach to argument structure*. Chicago: Chicago Univ. Press.

Goldberg, Adele. 2006. *Constructions at work*. Oxford: OUP.

Gries, Stefan. Th. 2007. *Coll.Analysis 3.2a. A program for R for Windows 2.x.*

Gries, Stefan, Beate Hampe and Doris Schönefeld. 2010. Converging evidence II: more on the association of verbs and constructions. In: Sally Rice and John Newman (eds.), *Empirical and experimental methods in cognitive/functional research*, 59–72. Stanford, CA: CSLI.

Gries, Stefan and Anatol Stefanowitsch. 2004. Extending collostructional analysis. A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9/1: 97–129.

Hacquard, Valentine. 2006. *Aspects of modality*. Ph.D. Thesis, MIT.

Hacquard, Valentine. 2009. On the interaction between aspect and modal auxiliaries. *Linguistics and Philosophy* 32: 279–312.

Hacquard, Valentine Actuality entailments. To appear in Lisa Matthewson, Cécile Meier, Hotze Rullmann and Thomas E. Zimmermann (eds.), *The Companion to Semantics*. Hoboken: Wiley.

Hilpert, Martin. 2016. Change in modal meanings. Another look at the shifting collocates of *may*. *Constructions and Frames* 8/1: 66–85.

Homer, Vincent. 2011. French modals and perfective: a case of aspectual coercion. In Mary Byram Washburn, Katherine McKinney-Bock, Erika Varis, Ann Sawyer and Barbara Tomaszewicz (eds.), *Proceedings of the 28th West Coast Conference on Formal Linguistics,* 106–114. Somerville, Mass.: Cascadilla.

Hunston, Susan. 2015. Lexical grammar. In Douglas Biber and Randi Reppen (eds.), *The Cambridge Handbook of English Corpus Linguistics*, 201–215. Cambridge: CUP.

Hunston, Susan and Gill Francis. 2000. *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

Koch, Peter and Wulf Oesterreicher. ²2011. *Gesprochene Sprache in der Romania*. Berlin: De Gruyter.

Labeau, Emmanuelle. 2002. L'unité de l'imparfait. Vues théoriques et perspectives pour les apprenants du français langue étrangère. *Travaux de linguistique* 45/2: 157–184.

Labeau, Emmanuelle. 2004. Les temps du compte rendu sportif francophone. *Journal of French Language Studies* 14/2: 129–148.

Labeau, Emmanuelle. 2006. French television talk: what tenses for past time? *International Journal of Corpus Linguistics* 11/1: 1–28.

Labeau, Emmanuelle. 2007. Et un, ou deux, ou trois? Les temps-champions du compte rendu sportif depuis 1950. In Emmanuelle Labeau, Carl Vetters and Patrick Caudal (eds.), *Sémantique et Diachronie du système verbal français*, 203–221. Amsterdam: Rodopi (= Cahiers Chronos 16).

Labeau, Emmanuelle. 2013. Les temps du compte rendu sportif francophone. *Journal of French Language Studies* 14/2: 129–148.

Laca, Brenda. 2012. On modal tenses and tensed modals. In Chiyo Nishida and Cinzia Russi (eds.), *Building a bridge between linguistic communities of the Old and the New World. Current research in tense, aspect, mood and modality*, 163–198. Amsterdam: Rodopi (= Cahiers Chronos 25).

Louw, Bill. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Mark Baker, Gill Francis and Elena

Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair*, 157–175. Amsterdam: John Benjamins.

Mari, Alda and Fabienne Martin, 2007. Tense, abilities, and actuality entailment. 151–156. Proceedings of the Amsterdam Colloquium.

Martin, Fabienne. 2009. Epistemic modals in the past. In Janine Berns, Haike Jacobs and Tobias Scheer (eds.), *Romance Languages and Linguistic Theory 2009*, 185–202. Amsterdam: Benjamins.

Michaelis, Laura. 2011. Stative by construction. *Language* 49/6: 1359–1399.

Nuyts, Jan. 2016. Analyses of the modal meanings. In Jan Nuyts and Johan van der Auwera (eds.), *The Oxford Handbook of Modality and Mood*, 31–49. Oxford: OUP.

Römer, Ute. 2009. The inseparability of lexis and grammar. Corpus linguistic perspectives. *Annual Review of Cognitive Linguistics* 7: 140–162.

Rubio Vallejo, David. 2017. Actuality effects as conversational implicatures. *Journal of Pragmatics* 112: 44–67.

Siepmann, Dirk. 2015. Dictionaries and spoken language: a corpus-based review of French dictionaries. *International Journal of Lexicography* 28/2: 139–168.

Siepmann, Dirk and Christoph Bürgel. 2015. L'élaboration d'une grammaire pédagogique à partir de corpus: l'exemple du subjonctif. In Thomas Tinnefeld (ed.), *Grammatikographie und Didaktische Grammatik – gestern, heute, morgen. Gedenkschrift für Hartmut Kleineidam anlässlich seines 75. Geburtstages*, 159–185. Saarbrücken: htw saar.

Siepmann, Dirk and Christoph Bürgel. 2016. Das *Corpus de référence du français contemporain* (CRFC) und sein Einsatz in der Grammatikographie am Beispiel des Präpositionsgebrauchs. In: Dirk Siepmann and Christoph Bürgel (eds.), *Sprachwissenschaft und Fremdsprachendidaktik. Zum Verhältnis von sprachlichen Mitteln und Kompetenzentwicklung*, 143–162. Baltmannsweiler: Schneider Hohengehren.

Siepmann, Dirk, Christoph Bürgel and Sascha Diwersy. 2017. The Corpus de référence du français contemporain (CRFC) as the first genre-diverse mega-corpus of French. *International Journal of Lexicography* 30/1, 63–84.

Stubbs, Michael. 2016. Corpus semantics. In Nick Riemer (ed.), *The Routledge Handbook of Semantics*, 106–121. London: Routledge.

Waugh, Linda R. and Monique Monville-Burston. 1986. Aspect and Discourse Function: The French Simple Past in Newspaper Usage. *Language* 62/4: 846–877.

Weinrich, Harald. 1982. *Textgrammatik der französischen Sprache*. Stuttgart: Klett.

Oscar Garcia-Marchena

Polar Verbless Clauses and Gapping Subordination in Spanish

Abstract Polar verbless clauses and gapping seem to differ in their capacity to be embedded. While polar verbless clauses can be easily subordinated, gapping constructions are traditionally considered main clause phenomena restricted to root contexts. Nevertheless, some languages, like Farsi, Rumanian and Spanish seem to allow gapping embedding with some particular predicates. This paper provides corpus data which show the extent of the capacity of subordination of Spanish gapping constructions, and their differences to the less restricted polar verbless clauses: on the one hand, gapping, like other fragments, can be embedded by verbal and non-verbal epistemic predicates. On the other hand, polar verbless clauses can be subordinated to these predicates, but are not restricted to them. They are much more frequently embedded, as can be seen by their distribution in the different genres of the CORLEC corpus.

Keywords Polar verbless clauses, gapping, subordination, fragments, ellipsis, embedding

1 Embedded polar verbless clauses

1.1 Introduction to embedded polar verbless clauses in Spanish

English polar verbless clauses can be defined as structures headed by a pro-clause, such as the polarity adverbs *yes* and *no*. It has been argued that polar clauses are not cases of ellipsis, but verbless clauses, because the pro-clause is anaphoric to a whole phrastic content of the type *message,* as is defined by Ginzburg and Sag (2000). In contrast, gappings are constructions where two clauses are coordinated and the verbal head of the second is elliptical. Also, in gapping, the remnant of the ellipsis is composed by two different phrases: the subject noun phrase and a phrase of the verb phrase. In this way, the elided verb leaves a gap between the two phrases.

Both polar verbless clauses and gapping constructions seem to be restricted to root sentences, as has been pointed out in several works, like Carlson (2001) and Merchant (2013), as is illustrated by (1a) and (1b) respectively. Nevertheless, Spanish seems to differ from English in this point, since it seems to accept these subordinations, as is suggested by the examples of Jimenez Julia (1995) (2):

(1) a. *John won't come to the party but I think Anne yes.
b. *John will have caviar, although others beans.

(2) a. Francisco quiere estudiar en la Universidad C, y creo que Javier en la A.
'Francisco wants to study in the Universidad C and I think Javier in the A.'
b. Andrés estudia Filología en Santiago, y me han dicho que Manolo no.
Andres studies Philology in Santiago and I have been told that Manolo not.'

Jimenez Julia's claim is supported by corpus data. In fact, the CORLEC corpus of contemporary oral Spanish (Marcos Marin 1992) provides evidence which supports this claim, showing that gapping and polar fragments are indeed frequent in subordination in oral Spanish. The CORLEC corpus is composed of 63,000 utterances and classified by genre, such as *university lessons, high school lessons, TV news, informal conversation, broadcasting of sports events*, etc. We have classified these genres as either monologic or dialogic, depending on whether the utterances in a specific genre are generally produced by one or by several speakers.

In this corpus we find 543 cases of subordinated fragments, with a higher frequency in dialogic genres than in monologic genres (390 vs. 153 examples), and particularly in the genre *informal conversation* (183 examples). Examples of gapping in subordination are less frequent, but corpus data show that they are nevertheless employed in both kinds of genres: 43 cases in dialogic and 26 in monologic genres.

This article aims to present the syntactic diversity and corpus frequency of subordinated gapping and polar fragments in Spanish, as shown by the data in the CORLEC corpus. It will focus on the syntactic structures that can be found, the part of speech at the head of gapping fragments, and the syntactic type and illocutionary value of the clause where the fragment is embedded.

As said before, polar verbless clauses are headed by the polarity adverbs *yes / no*, which constitute pro-sentences that are anaphoric to a clausal content previously uttered (3) or a part of it (4). These structures are therefore non-elliptical, since their content is either present or recovered by anaphora. They are also syntactically complete, since they form whole syntactic structures where all elements meet their sub-categorisation requirements. These properties define them

as verbless clauses, and distinguish them from fragments such as (1b) and (2a), which do have elliptic content.

(3) A: -¿Hay o no hay? B: -Yo creo que sí. (EDU 018A)
A: -'Is there some or not?' B: -I think that yes. (I think there is)
(4) El problema es que él no puede aparcar tal y como está y yo sí. (CONV 119A)
'The problem is that he can't park as it is and I yes.' (and I can)

Some works, like Laka (1990) and Kramer & Rawling (2009), note that they cannot be embedded in English with an overt complementizer[1], as in the example **I suspect that {yes / no}.* Other works, (such as Sailor 2012) observe that this is not a general property of pro-sentences, since polar verbless clauses can indeed be embedded in other languages, like French (5a), Spanish (5b), Catalonian (5c), Hebrew (5d) and Russian (5e):

(5) a. Je pense que {oui / non} Lit.: 'I think that {yes / no}'
b. Creo que {sí / no} id.
c. Crec que {sí / no} id.
d. Ani xoshev she {ken / lo} id.
e. Ja dumaju cto {da / net} id.

1.2 Syntactic diversity of embedded polar verbless clauses in the corpus CORLEC

The corpus of oral Spanish CORLEC (Corpus oral de Referencia de la Lengua Española Contemporánea (Marcos Marín 1992)) supplies quite a number of examples of embedded polar verbless clauses. It is composed of 1,078,780 words, distributed in 63,291 utterances and classified by genres, which we have grouped as either dialogic or monologic. Among these, we find 734 subordinated verbless utterances, distributed in fragments (20,84 %, 153 items) and embedded verbless clauses (79,16 %, 581 items). Most of the embedded verbless clauses found in the corpus are polar verbless clauses (543 items), which shows they are often used in embedding contexts.

1 Examples without complementizer can be found (*I think yes*), and can be analyzed as transcriptions of two juxtaposed units (*I think, yes*). Also, as noted by an anonymous reviewer, relevant cases of subordinated *yes* and *no* can be heard in pidgin varieties of English.

Their distribution in the corpus shows that embedded polar verbless clauses are generally more frequent in dialogic contexts (390 items) than in monologic ones (153 items). Also, among the monologic genres, they are extremely frequent in the genre *instructions* (62 items), since this genre shares many properties with dialogic genres: in instructions, speakers do not interact with listeners, but nonetheless they ask them rhetorical questions like (6a) to ensure they follow the conversation. Many of these rhetorical questions display the subordination of polar verbless clauses, as in (6b). In the same way, polar verbless clauses occur infrequently in dialogic genres composed by short sentences and little subordination like *administration, sports* and *publicity*. These distributions are illustrated in table (1).

(6) a. -¿Ves? 'You see?' (LUD 002A)
b. -¿A que sí? 'Isn't it?' (EDU 013A)

Table 1: Genres and frequencies of polar verbless clauses.

GENRE	**ID**	**Freq**	**GENRE**	**ID**	**Freq**	**TOTAL**
Dialogic			**Monologic**			
Administrative	Amin	0	Religion	Rel	5	
Sport	Dep	9	Instructions	Ins	62	
Publicity	Pub	15	Documentary	Doc	9	
Debate	Deb	42	University	Hum	11	
High School	Edu	20	Science	Cie	6	
Games	Lud	33	Law	Jur	14	
Interviews	Ent	88	Politics	Pol	10	
Informal	Conv	183	Technique	Tec	18	
			News	Not	18	
Sub-total		**390**			**153**	**543**
Sub-average		**48,75**			**60,33**	

This data shows that polar verbless clauses are the most frequent verbless structure found in subordination. A closer look at the examples reveals a great deal of syntactic diversity, as we find different syntactic structures: they can be composed of only a head (7a), of a structure head-complement (7b) or of a head-adjunct (7c). The head can also appear in the left-periphery of the clause together with a dislocated phrase, forming a head-peripheric structure, as in (7d). We can indeed note in (7d) that the demonstrative pronoun *ese* 'that' preceding the polar adverb *sí* is neither the subject nor specifier of the polar head, but only a

dislocated (peripheric) pronoun. We find not only structures composed by two phrases, but also by three or more, as in (7e).

(7) a. Les preguntamos si podíamos traer invitados y nos dijeron que no. (CONV 042B)
'We asked them if we could bring any guests and they said that no.'
b. Seguro que sí que viene a decirnos algo. (CONV 152A)
'Sure that yes that he comes to tell us something.'
c. Me gusta comerlo y tal pero todos los días no. (CONV 061A)
'I like to eat it and stuff but every day not.'
d. No llegan nunca a tener éxito, porque ese sí que es un precio demasiado alto que tiene que pagar la sociedad. (DEB 17)
'They never get to be successful, because that one yes that it is a price too high that society has to pay'
e. El juez en un momento determinado sí. (JUR 003A)
'The judge in a given moment, yes.'

Polar verbless clauses can be embedded in clauses of different syntactic types: declarative, without any formal marker of syntactic type (8a) or interrogative, with an interrogative word like *cómo* in (8d). Nevertheless, we do not find examples of either desiderative or exclamative types, and the latter seem to be only marginally acceptable, as shown in (8e). Declarative clauses can have different illocutionary values, since they can be used to convey an assertion (8a), an exclamation (8b) or an question (8c):

(8) a. Todos queremos resolver esto, ¡claro que sí! (POL 010A)
'We all want to solve this, of course!'
b. ¡Vaya que sí!
'Of course that yes!' (yes indeed!)
c. El caviar persa, lo mejor es tomárselo sólo, ¿verdad que sí? (CONV 021A)
'Persian caviar, it is better to have it alone, true that yes?'
d. A: -Yo no estaba. B: -¿Cómo que no? (CONV 029B)
A: -I wasn't there. B: -How that not? (=Sorry?)
e. ?¡Qué suerte que no! 'How lucky that not' ('Luckily not')

Nevertheless, these different syntactic configurations are not all equally frequent. Most embedded polar verbless clauses are composed of only a head (494 cases), although the polar head can also have a complement (28 cases), an adjunct (9 cases) or a dislocated element in the left periphery of the clause (12 cases). Similarly, most of them are declarative assertive (522 cases), although they can have a questioning value (12 cases) or an interrogative type (10 cases).

1.3 Types of subordinators

Polar verbless clauses can also be embedded by a variety of heads and constitute either their complement or their adjunct. In this way, they can be adjuncts to a noun (9abc) or to a verb (10), expressing a variety of semantic relations: cause (10a), condition (10bd) and time (10c). Interestingly, the embedded polar verbless clause is not always interpreted as a verbal adjunct, as in (10c). Sometimes it is interpreted as an illocutionary adjunct, like in (10abd); in (10a), the adjunct seems to be a cause of the illocutionary act of saying, being interpreted as *'I say that because…'* Similarly, in (10bd), the condition does not seem to rely on the predicate, but on the illocutionary act of committing, so for (10b), rather than 'If I don't call you and if we cannot meet…' the interpretation seems to be: 'Instead of committing to this…'

Polar verbless clauses can also constitute the complement of a verb (11a) or of an adjective (11bc). We can note that all cases of polar verbless clauses embedded as a complement are complements of epistemic predicates like the verbs *decir* 'say', *responder* 'answer', *parecer* 'seem', *imaginar* 'imagine', *puede ser* 'maybe' and *temer* 'fear', and like the adjectives *cierto* 'certain' *seguro* 'sure' and *claro 'clearly'*.

(9) a. Hay chicos que sí que saben comportarse. (CONV 023A)
'There are children that (yes that they) can behave.'
b. Las instrucciones normalmente suele venir, pero a veces hay algunos que no. (DEB 023A)
'The instructions usually come with it but sometimes there are some that not.'
c. (…) en vez de éstas de abrir y cerrar, que sí que están bien, pero es innecesario. (CONV 11A)
'instead of this ones to open and close, that yes are good, but it is unnecessary.'

(10) a. La gente exterioriza más su sociabilidad, porque eso sí que lo puedo decir. (CONV 006A)
'People reveal more their sociability, because that yes that I can tell you.'
b. Te llamo yo, sí, y podemos quedar, o si no, espera un momento. (CONV 000A)
'I will call you, yes, and we can meet or, if not, wait a moment.'
c. Muchas veces me dan las doce o la una de la noche, y cuando no, pues te levantas a las tres de la mañana. (POL 007A)
'Often I stay awake until midnight or 1am, and when not, you would wake up at 3 am.

d. Esto sería una cierta sorpresa (…), como no, se quedaría sin representación parlamentaria. (NOT 019A)
'This would be a certain surprise (…), how not, he would lose his seat in the Parliament.'

(11) a. A: -¿Qué significa, que no vale el texto entero?
B: -Pues me temo que no. (DEV 014A)
A: -'What does it mean, that the whole text is not good?'
B: -'I fear that not.'
b. Es cierto que ahora sí que existe un comité de las regiones. (POL 006A)
'It is true that now yes that it exists a committee for regions.'
c. A: -¿Es que no hay variación de temperatura?
B: -Claro que sí. (CIE 006A)
A: -'Is there not a variation of temperature?'
B: -'Clear that yes (=of course)'

They can also appear as complements of a noun (12), of an adverb (13) or of a prepositional phrase (14). Furthermore, they can also be embedded to polar verbless clauses in root position, as complements (15a) or as adjuncts (15b).

(12) a. ¿Y es cierto? Porque corre el rumor de que sí. (ENT 040A)
'And is it true? Because I have heard the rumor that yes.'
b. A: -¿No le no le da a usted miedo (…)?
B: -Yo tengo confianza en que no. (ENT 012A)
A: -'Don't you fear that?' B: 'I have faith that not.'

(13) a. Si me dieran alternativas, para modificarlo, naturalmente que sí. (ENT 059A)
'If I were given alternatives to modify it, naturally that yes.'
(= of course I would)
b. ¡Ojalá que no! (ENT 064A)
'I-wish that not' (= I wish it won't)

(14) a. A: -¿Y usted cree, como escritor, que se podría hacer?
B: Sí: Por supuesto que sí. (POL 010A)
A: And you, as a writer, do you think it could be done?'
B: 'Yes. Of course that yes.'
b. A: -Un percance realmente singular.
B: -Sí, desde luego que sí. (NOT 012A)
A: 'An incident really particular'.
B: 'Yes, of course that yes.'

(15) a. Usted ha mandado su propio signo (…) y a eso sí que no se lleva los nueve millones. (PUB 033A)
'You have sent your own sign (…) and for that yes that you don't get nine millions.'
b. No, porque sí que tengo que conducir. (CONV 031B)
'Not, because yes that I have to drive.'

1.4 The subordination of Spanish polar verbless clauses: Conclusions

These corpus data show a number of properties of Spanish polar verbless clauses that can be generalised: they can be easily embedded, as both adjuncts and complements. Firstly, as complements of nouns, they provide the content of one of the arguments of the noun (12). Secondly, as complements of verbs, adjectives, adverbs and prepositional phrases, they have epistemic meanings (11), (13), (14). Thirdly, as adjuncts, they focalise the polarity adverb that can be cataphoric to a clausal content syntactically realized as its complement (9ac) (10a) or anaphoric to a previous content, being therefore placed in focus (final) position (9b) (10bcd).

Therefore, embedded polar verbless clauses provide an argument where polarity is focalised, and can be found in two different contexts: on the one hand, they can be complements of predicative heads with an epistemic content (verbs, adjectives, adverbs or prepositional phrases), and on the other hand, they can be complements of nouns or adjuncts of noun or verbs. It seems that in the last case, the embedded polar verbless clause can constitute an adjunct of the illocutionary act instead of an adjunct of the verb (10abd).

2 Subordinated Gapping in Spanish

2.1 Introduction to Embedded Gapping in Spanish

Gapping, the construction found in the second conjunct of a coordination where the verb is elided (16a), is a major subject in the literature on ellipsis. It is traditionally accepted that it cannot be embedded in English (Niejt 1979, Hankamer 1979, Johnson 2014) (16b). Similarly, the gapping antecedent cannot be embedded either (16c). In spite of this, the corpus CORLEC of contemporary oral Spanish offers some examples of subordinated gapping (17a) and of cases where the clause that contains the antecedent of the gapped constituent is subordinated (17b).

(16) a. Some ate beans and others, rice.
b. *Alfonse stole the emeralds, and I think that Mugsy the pearls.
c. *I think that Alfonse stole the emeralds, and Mugsy the pearls.

(17) a. Pero el chico la ama y dicen que ella a él. (CONV 033A)
'But the boy loves her and they say she him.'
b. Parece que el tío se fue a su casa y ella a la suya. (CONV 009A)
'It seems that the guy went to his house and her to hers.'

These examples of embedded gapping could be interpreted as the result of dysfluent productions, such as overlappings, grammatical errors or hesitations. Nevertheless, the CORLEC corpus is annotated for these types of dysfluencies. This suggests that embedding gappings are not the result of disfluency, but of a different syntactic configuration available in Spanish. This seems to contradict Johnson (2014), who, following Niejt (1979), states that the constraint that gapping cannot be embedded is a structural constraint of language, with a few exceptions: firstly, gapping can be a non-initial conjunct in an embedded clause containing a coordination if its antecedent is in a preceding conjunct (18a); secondly, gapping can be embedded if the antecedent and the gap are subordinated by an infinitive (18bc). Thirdly, gapping can be subordinated if the remnant is a wh-phrase (18d) (Niejt 1979).

(18) a. Jerome wishes that [Julie had bought a dress and Jennifer a pair of shoes.]
b. John tried to put his car in the garage and his bike in the barn.
c. John seems to be happy and Mary unhappy
d. Charles may decide which boys are coming along and Max which girls.

Johnson (2014: 7) describes this constraint of gapping as the *No Embedding Constraint*, formulated as follows:

> "Let A and B be conjoined or disjoined phrases, and β be the string elided in B whose antecedent is α in A. Then α and β must contain the highest verb in A and B."

2.2 Gapping subordination in other languages

This constraint has nevertheless been recently questioned for Farsi by Farudi (2013), who furnishes data where gaps occur in embedded contexts (19a). According to this work, Farsi also allows the antecedent of the gap to be in an embedded clause (19b). Furthermore, both the gap and its antecedent can be embedded (19c).

(19) a. maman chai xord va fekr mi-kon-am baba qahve.
'Mother drank tea and I think father ___ coffee.'
b. Fekr mi-kon-am ke Nasrim gormeh sabzi-ro dorost kard va man adas polow-ro.
'I think that Nasrin made spinach stew and I lentil rice.'
c. Ajib nist ke Râdmehr mâhiro xorde vali ajibe ke Ânâhitâ gushtro.
'It's not unusual that Rodmehr ate fish, but it's strange that Anahita meat.'
d. mujhe lag-taa hai ki mummi=ne caai pii thii lekin mujhe nahiiN lag-taa ki papa=ne coffee. 'I think that mother drank tea, but I don't think that father _____ coffee.'

Interestingly, in these examples, gapping is embedded under an epistemic predicate like *to think, to know, to be possible, to hear, to be strange* and *to be unusual*. Nevertheless, Farudi (2013) argues that these verbs are not parenthetical; if they were, they would not be able to establish syntactic or semantic relationships with the embedded clause, such as negation, wheareas negation is possible (19d). In conclusion, for Farudi (2013), these data seem to provide evidence that the restrictions on gapping embedding are not a universal property of gapping; rather, they seem to be at work in only some languages.

Gapping embedding also seems possible in other languages. In Romanian, Bîlbiie (to appear) shows that verbs which allow gapping embedding are a particular class of verbs, with a particular syntactic behaviour. They express an epistemic content, especially in the first person (20). In this way, these verbs with epistemic content have particular properties which distinguish them from other verbs. They have received different analysis, as "weak verbs" (Blanche-Benveniste & Willems 2007), "grafts" (van Riemsdijk 2006) or "hedges" (Lakoff 1973).

(20) a. Andrei a luat cartea şi cred că Marga atlasul.
'Andrei has taken the book and I think that Maria the atlas'
b. Ion are trei copii şi pare-se că Maria doar unul.
'Ion has three children and it seems that Maria only one.'

2.3 Gapping subordination in Spanish

The corpus data from the CORLEC suggest that in Spanish, like in Farsi, and more than in Rumanian, gapping can be embedded in various contexts, such as with an impersonal form (21a). Also, Spanish, like Farsi, respects island constraints, like the relative clause and indirect question constraints (21bc).

(21) a. Pero el chico la ama y dicen que ella a él. (CONV 033A)
'But the boy loves her and they say she him.'
b. *Luis quiere ir a sitios que tengan playa y Sara prefiere sitios que montaña.
'Luis wants to go to places with beach and Sara prefers places that mountain.'
c. *Tu no sabes quién compró el vino y yo no sé quién el pan
'You don't know who bought the wine and I don't know who the bread.'

Gapping can also be analysed as a particular construction where a fragment with an unheaded structure is coordinated to a clause with a verbal head. These same fragments can appear in other contexts, such as answers (22a), and they can also be embedded by epistemic verbs like *to say* (22b). We even find in the corpus examples of answers (or reactions) composed of a coordination of two fragments where only one of them is embedded by an epistemic verb and the antecedent of both is in the previous utterance (22c)

(22) a. A: -¿Quién iba agarrado de quién?
B: -Vicky de un niño pequeñito. (CONV 112B)
A: -'Who was holding who? Vicky a little child.'
b. A: -¿Sabéis en qué situación podéis quedar los trescientos trabajadores del independiente?
B: -Pues por aquí dice una compañera que la mayoría en la calle. (DOC 010A)
A: 'Do you know which situation can the three hundred independent workers expect?'
B: -'Here a colleague says that the majority in the street.'
c. A: -Esperemos que hagan un poco más de las cuarenta mil, que es más o menos la media habitual de las taquillas.
B: -Pues el otro día, Juanjo, _____ treinta y dos mil contra el Vicálvaro, y hoy me da que _____ ni la mitad, vamos. (DEP 013A)
A: -'Let's hope they earn more than forty thousand, which is more or less the usual average in the ticket window.'
B: 'The other day Juanjo thirty-two thousand against Vicalvaro, and today I think not even half of them, let's see.'

This contrast suggests that epistemic predicates can embed not only gapping constructions, but also fragments such as those found in answers or reactions. These data allow us to draw two main conclusions: firstly, it seems that a variety of epistemic heads (verbs or other predicative part-of-speech) can embed gapping constructions. In this way, the analysis of these predicates as weak verbs

(Blanche-Benveniste & Willems, 2007) must be extended to non-verbal parts-of-speech. Secondly, it seems that these predicates can not only embed gapping, but also other types of fragments (22) or even verbless utterances.

3 Conclusions

The corpus data presented here on Spanish gapping and polar verbless clauses, and the contrast with data from other languages, allow us to draw a number of conclusions. Firstly, it seems that weak (epistemic) heads are not limited to verbs, but may also extend to non-verbal predicative heads. Secondly, these weak heads behave differently in different languages, since some of them allow more embedding than others. Thirdly, Johnson's (2014) embedding constraint should be enriched with an account of weak verbs to deal with cross-linguistic variation.

In this way, some elliptical constructions like gapping or fragments seem to constitute root phenomena, excluded in embedded contexts. Nevertheless, weak predicates seem to constitute an exception to this. This unorthodox embedding has a particular behaviour that has been described in several works (de Cuba & MacDonald 2013, Fernández-Sánchez 2016): syntactically, they constitute fully integrated predicates, as showed by Farudi (2013). Semantically, they provide a content which is not the main content of the utterance. Indeed, the main content is supplied by the embedded clause, whereas the weak predicate is limited to expressing an epistemic modality. Pragmatically, the embedded predicate in the root clause has the discursive function of an evidential marker, making explicit the speakers' reason for asserting the content of the complement.

In some languages, evidentiality is morphologically marked, as logophoric pronouns (Weir 2014: 242), or as bound morphemes (Aikhenvald 2004). This capacity for embedding epistemic predicates seems to constitute a syntactic means of encoding evidentiality. Languages seem to differ in the extent to which they allow embedding of weak epistemic predicates, leading to cross-linguistic differences.

Finally, we have observed that gapping embedding is restricted to weak epistemic predicates, whereas polar verbless clauses can also be subordinated to other heads, as adjuncts or as noun complements. This difference shows that polar verbless clauses are less restricted. If they can be embedded by weak heads, like gapping and other fragments, they can also share with verbal clauses the capacity of being subordinated as adjuncts or noun complements. These properties mark a direction for future research: what are the properties of other types of verbless clauses and fragments regarding subordination? Can all fragments be embedded only by weak epistemic heads? Are verbless clauses less restricted in this regard?

References

Aikhenvald, Alexandra Y. 2004. *Evidentiality*. Oxford: Oxford University Press.

Bîlbiie (to appear) Grammaire des constructions elliptiques : une étude comparative des phrases sans verbe en roumain et en français. Ph. D. thesis, Université Paris Diderot – Paris 7.

Blanche-Benveniste, Claire and Dominique Willems. 2007. "Un nouveau regard sur les verbes faibles." *Bulletin de la Société Linguistique de Paris* 102/1: 217–254.

Carlson, Katy, Charles Clifton Jr., and Lyn Frazier. 2001. Prosodic boundaries in adjunct attachment. *Journal of Memory and Language* 45: 58–81.

De Cuba, C. and J.E. MacDonald. 2013. On the referential status of embedded polarity answers in Spanish. In Selected proceedings of the 16th Hispanic Linguistics Symposium (pp. 312–23).

Everaert, M. and H. van Riemsdijk. 2006. The Blackwell Companion to Syntax, Volume I. Blackwell.

Farudi, A. 2013. Gapping in Farsi. A cross-linguistic investigation. PhD dissertation. University of Massachusetts Amherst.

Fernández, R. and J. Ginzburg. 2002. Non-sentential utterances: A corpus study. *Traitement Automatique des Languages: Dialogue* 43 (2): 13–42.

Fernández-Sánchez, J. 2016. Topics at the left edge of infinitive clauses in Spanish and Catalan. *Borealis–An International Journal of Hispanic Linguistics*, 5(2): 111–134.

Ginzburg, J. 2012. *The Interactive Stance*. Oxford University Press.

Ginzburg, J. and I. Sag. 2000. *Interrogative Investigations: the Form, Meaning and Use of English Interrogatives*. Stanford : CSLI.

Hankamer, J. 1973 Unacceptable ambiguity. *Linguistic Inquiry* 4:17–68.

Jiménez Juliá, T. 1995. La coordinación en español. Aspectos teóricos y descriptivos, Universidade de Santiago de Compostela, in Verba, 39.

Johnson, K. 2009. Gapping is not (VP-) ellipsis. *Linguistic Inquiry* 40(2), 289–328.

Johnson, Kyle. 2014. Gapping. Ms., University of Massachusetts Amherst.

Johnson, K. 2005. Gapping. In: Martin Everaert and Henk van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, Volume 1, 407–435. Oxford: Blackwell.

Kramer, Ruth, and Kyle Rawlins. 2009. Polarity particles: an ellipsis account. In: *Proceedings of NELS 39*. University of Massachusetts.

Laka, Itziar. 1990. Negation in syntax: On the nature of functional categories and projections. Doctoral Dissertation, MIT.

Lakoff, R. 1973. Language and Woman's Place. *Language in Society*, Vol. 2, No. 1: 45–80.

Marcos-Marín, F. 1992. Corpus de referencia de la lengua española contemporánea: Corpus oral peninsular. Technical report, Universidad Autónoma de Madrid.

Merchant, J. 2013. *Ellipsis: A Survey of Analytical Approaches*. University of Chicago.

Muysken, P. and T. Veenstra. 2006. Serial verbs. In: M. Everaert and H. van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, Volume IV, 234–270. Oxford: Blackwell.

Neijt, A. 1979. *Gapping: A Contribution to Sentence Grammar*. Dordrecht, Foris Publications.

Riemsdijk, H. 2006. Free relatives. In: M. Everaert and H. van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, 332–382. Oxford: Blackwell.

Sailor, Craig. 2012. Remarks on replies: On emphatic polarity in English. Presented at the 2012 LSA Annual Meeting.

Weir, A., 2014. Fragments and clausal ellipsis. Doctoral dissertation, University of Massachusetts, Amherst.

Laura Becker

Aspectuality in Hungarian, German, and Slavic. A Parallel Corpus Study

Abstract The present paper compares verbal prefixation in Hungarian and German with the expression of aspect in Russian and Czech based on parallel movie subtitles. In order to account for interactions between lexical (actional) properties and aspect, four classes of verbs are considered: relative-statives, activities, gradual-terminatives, and total terminatives. The other factors examined with respect to their relation to aspect marking are: presence of a prefix, presence of a suffix, tense, mood, negation, transitivity (presence of an accusative argument). Results show that Hungarian patterns with Slavic for relative-statives and total-terminatives, while it is similar to German for activities and gradual-terminatives. This hybrid behavior of Hungarian is confirmed by the importance of the factors: in both Slavic languages, the presence of the prefix has the greatest influence on the aspect choice, followed by actionality, tense, and mood. In Hungarian and German, however, actionality is the most relevant factor; therefore, despite many similarities between Hungarian and Slavic, aspect cannot be viewed as grammatical in Hungarian.

Keywords Aspect, Slavic, Hungarian, prefix, parallel corpus

1 Introduction

Aspectuality in Slavic is a well-known and widely discussed topic, as it has been argued to be expressed grammatically and, at the same time, involve derivation (e. g. Dahl 1985; Lehmann 1999). The latter is dominantly expressed by verbal prefixes[1], which are found in a very similar form and function in Hungarian as

1 It is not uncontroversial to assume that prefixation is part of aspect formation proper in Slavic. Isačenko (1960), for instance, argued that real aspectual pairs are only those formed by suffixation. In this paper, I follow the more liberal tradition assuming that at least some prefixes are able to function as proper perfectivizers in combination with certain verbs.

well. For Hungarian, different analyses of the verbal prefixes have been proposed. Some authors argued that they are perfectivizers (e. g. Soltész 1959; Piñón 1995; Kiefer 2006; É. Kiss 2006), while others attributed only a telicizing function to prefixes, delimiting the situation due to the lexical content of the prefix in interaction with the verbal semantics (e. g. Dahl 1985; Eördögh 1986; Csató 1994). Therefore, it is still not clear to what extent aspectuality is grammaticalized in Hungarian. The aim of the present study is to address this issue empirically and to determine to what extent the presence of verbal prefixes and the expression of aspectuality are correlated in Hungarian.

To do so, verbal prefixation in Hungarian will be compared with that of German, as well as with the expression of aspectuality in Russian and Czech. The latter two languages will serve as "aspect" base line and ensure that potential inner-Slavic variation between East- and West-Slavic is accounted for (Dickey 2000; Wiemer 2008). German will be considered for its formally similar system of verbal prefixation, which is not involved in the marking of aspect.

The corpus used consists of parallel movie subtitles from the four languages. By using parallel texts, semantics and pragmatics can be controlled for, which makes aspectual marking directly comparable across languages. Also, the similarity of form-function mapping of aspect in the different languages can be measured, so that Hungarian verbal prefixation and the expression of aspect can be situated between German (no aspect marking) and Czech/Russian (Slavic aspect).

2 Aspect (in Slavic)

2.1 General remarks on aspectuality

There is a general consensus that aspectuality, especially with respect to Slavic, is primarily a matter of 'boundaries'", meaning that we deal with temporal boundaries of situations (Sasse 2001). Aspectuality can be coded grammatically. In that case, we speak of aspect, which must represent a grammatical category (Dahl 1985:23; Lehmann 1999:218). In order to constitute a grammatical category, the following (idealized) criteria should hold: (i) aspectual values must be abstract and not concrete, (ii) aspect must affect the entire verbal system, and, with respect to the Slavic aspect, (iii) it must feature a binary opposition of imperfective and perfective values. Since the present paper addresses the Slavic aspect type, the following paragraphs will focus on the properties of the latter type only.

The perfective value marks situations as bound in time, its core functions cover the expression of sequences of situations and single events. The imperfective value, on the other hand, presents situations as unbound in time, and is typically used to denote parallel and repeated situations.

The most frequent formal pattern[2] to derive perfectives and imperfectives in Slavic begins with a simple verb that has been reinterpreted as imperfective. A perfective counterpart can be derived by prefixation from such a verb as *pisat'* 'write', e. g. *na-pisat'* 'write (pfv). Since, in some cases, the prefix might add lexical semantics to the verb meaning, we also find new lexemes derived by prefixation (e. g. *pere-pisat'* 'write anew' from *pisat'* 'write'). To form an imperfective counterpart of the perfective, lexically-modified verb, Slavic features suffixation to form "secondary imperfectives" such as *pere-pis-yvat'* 'write anew', which constitute lexical counterparts of the imperfective form.

The Slavic aspect system is highly intertwined with tense. What is formally a present tense perfective has been reinterpreted as future (with a few exceptions). Therefore, the perfective aspect is incompatible with the present tense meaning. Imperfectives, on the other hand, have developed an analytic future tense.

2.2 Aspect and actionality

The notion of aspect, denoting a grammatical phenomenon, is usually employed in opposition to aktionsart as lexical phenomenon.[3] For the present purposes, we will distinguish between aspect, i. e. externally set boundaries of a situation independent of inherent semantics of the verb, and actionality (cf. Tatevosov 2002), the latter marking telicity[4], the inherent boundaries of a situation dependent on the semantics of the verb. These two levels have to be distinguished from each other, since they can combine in the ways displayed in Table 1.

Table 1: Combination of aspect (terminativity) and actionality (telicity).

		telicity	
		telic	atelic
terminativity	perfective	*po-stroit'* 'build (up)'	*po-rabotat'* 'work for some time'
	imperfective	*na-xodit'* 'find'	*igrat'* 'play'

There are different proposals to integrate actional properties into the selection of aspectual values for Slavic (e. g. Breu 2000; Tatevosov 2002; Lehmann 2009).

2 Note that aspectual pairs can also be marked by other mechanisms: suffix opposition (*stučat'* "knock (ipf) vs. *stuknut'* "knock (pfv)") and suppletion (*brat'* "take (ipf)" vs. *vzjat'* "take (pfv)").

3 Another prominent approach that distinguishes between lexical (situation) and grammatical (viewpoint) aspect is found in Smith (1997).

4 For this use of the term telicity, also see Arkadiev (2015).

The present study adapts the classification of interactions between aspect and actionality from Breu (1994, 2000). The author distinguishes the following actional classes: (i) total-statives, (ii) relative-statives, (iii) activities, (iv) total-terminatives, (v) gradual-terminatives, (vi) inceptive-statives, and (vii) inchoatives. Section 4.2 will discuss the classes in detail.

3 Verbal derivation in Hungarian and German

3.1 Verbal prefixation and suffixation in Hungarian

Hungarian verbal prefixes have mostly been studied for their syntactic properties, as they are separable from the rest of the verb under certain syntactic, formal, or pragmatic conditions (e. g. É. Kiss 2006; Ladányi 2015). Therefore, they have also been referred to as verbal particles. For the sake of comparison, they are labeled as prefixes in the present paper.

Similar to verbal prefixes in many languages, most Hungarian prefixes originate from spatial expressions (Ladányi 2015). The most frequent ones are: *be* 'into', *ki* 'out', *fel* 'up', *le* 'down', *el* 'away', *meg* 'completely'[5].

A prominent function of prefixes, especially of *meg*, is to mark applicatives (1) and upgrade oblique arguments to direct objects (2):

(1) a. ajándékoz egy könyvet
give.as.present a book-ACC
'give a book as present'
(Hungarian)

b. **meg**-ajándékoz egy barát-ot
PFX:APPL-give.as.present a friend-ACC
'make a present to a friend'
(Hungarian)

(2) a. beszél a helyzet-ről[6]
talk the situation-DELAT
'talk about the situation'
(Hungarian)

5 This prefix originates from an expression for 'behind', but has lost its lexical semantics almost completely in the current language.

6 The glossing of examples follows the Leipzig Glossing Rules (https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf). Other less common abbreviations used are: DELAT = delative, PFX = prefix, SUPERESS = superessive.

b. **meg**-beszéli a helyzet-et
PFX:APPL-talk.DEF the situation-ACC
'address the situation'
(Hungarian)

The other main function is telicization. By adding a goal/delimitation to the verbal meaning, the prefixes telicize the denoted situation, as is shown in the example below:

(3) a. épít egy város-t
build a city-ACC
'build a city'
(Hungarian)
b. **fel**-épít egy város-t
PFX:UP-build a city-ACC
'build up a city'
(Hungarian)

What does not seem to be clear until now is whether those telicizing prefixes also perfectivize the situation, i. e. whether delimitation only operates on a lexical or on a more systematic, grammatical level. To illustrate this, two examples from the corpus are given in (4) and (5).

(4) Persze nem történt volna **meg**, [...][7]
of.course NEG happen.PST.3SG IRREAL **PFX**
'Of course, none of it would have happened'
(Hungarian, Frozen)

(5) Hogyan találta **meg**?
how find.PST.3SG.DEF **PFX**
'How did you find it?'
(Hungarian, Inception)

In the examples above, the verbs for 'happen', and 'find' are telic, which means that the function of the prefix cannot be to telicize the situation. Rather, they seem to point out and highlight the telic semantics of the verb. Whether this occurs on a more abstract and systematic level, which would be required to label it aspect, cannot be discussed based on these few examples alone, but is an empirical question and will be addressed in section 5.

7 This example, as well as the following ones except (6) and (7), are taken from the subtitle corpus. The English translations given are the original subtitle lines.

In addition to prefixes, Hungarian also features a few derivational suffixes on the verb that are somewhat involved in the expression of actionality, e. g. deriving frequentatives. However, those suffixes are lexically restricted and occur idiosyncratically. Therefore, they are not considered in the present study.

3.2 Verbal prefixation in German

Verbal particles in German (labeled prefixes in this paper) have also been addressed in previous research with respect to their syntactic status and semantic functions (e. g. Stiebels 1996, Lüdeling 2001); however, they are usually not associated with aspect. In combination with many verbs, prefixes in German add spatial orientation to the verbal semantics, as is shown in (6) below.

(6) **hinein**-legen
PFX:INTO-put
'put into'
(German)

Also, applicatives and upgrading of oblique arguments into direct object positions are marked by prefixes on the verb:

(7) a. ein Buch schenken
a.ACC book.ACC offer
'offer a book'
(German)
b. einen Freund **be**-schenken
a.ACC friend.ACC **PFX:APPL**-offer
'give a friend a present'
(German)

Prefixes can also be used to telicize situations, especially if the verbal semantics include an endpoint or limit of the situation that can but does not have to be reached in a given instance. In these cases, the prefix points to that endpoint and hence delimits the situation expressed. Examples (8) and (9) from the corpus below illustrate this:

(8) Wärm dich **auf**.
warm.IMP yourself.ACC **PFX:UP**
'Get warmed up.'
(German, Black Swan)

(9) Das Herz ist nicht leicht zu **ver**-ändern.
the.ACC heart.ACC is NEG easily to **PFX**-change
'The heart is not so easily changed.'
(German, Frozen)

The two simple verbs *wärmen* 'warm' and ändern 'change' refer to situation with no endpoint inherently implied. When combining with a prefix, the latter points to that endpoint so that the situation necessarily is presented as telic.

4 Methodology

4.1 Corpus and annotation

The corpus used for the present study includes subtitles from the movies Avatar, Black Swan, Frozen, Noah, and Inception (Levshina 2016). From these subtitles, the first 1000 sentences with different verbal lexemes which fulfilled certain requirements (see below) were extracted. Finally, 578 verbs in Russian, Czech, Hungarian, and German were manually annotated for the four languages, so that, in total, 2312 data points could be considered. The choice of tokens was not restricted to certain lexemes to avoid potential bias by particular lexemes. Also, no restriction on the verb classes was made to determine the frequency distribution of those classes is in natural usage (which proved to be fairly equally distributed). Crucial for the choice of tokens, however, was that the meanings of the verbs in the four languages were sufficiently similar.[8]

The predicates selected were annotated for the lexeme, actionality, aspect, presence of a prefix, presence of a suffix (only for Russian and Czech), negation, tense, mood, presence of an accusative object.

We will now address the annotation and values of the factors considered in more detail. As for the prefix, only the presence or absence was noted, independently of whether it derives a new lexeme and/or is no longer separable from the rest of the verb on the synchronic level (e. g. Russian ***u**bit'* 'kill', Czech ***na**jít* 'find', Hungarian ***be**fejez* 'end', and German ***er**zählen* 'tell').

It has also been noted in which cases it is the presence of the prefix that perfectivizes; for a verb like *sozdat'* "create (pfv)" the prefix (*soz-)* was counted

8 Although the texts are parallel in the four languages and are used to accompany the same movie scenes, the languages use other constructions, predicates, and sentence types in some contexts. Only those contexts with verbs of shared lexical semantics and the same participants were considered in this study. Note that the verbs in each language, even if sharing lexical meaning, do not necessarily belong to the same actional class.

in. However, it was not counted as perfectivizing prefix, since, synchronically, it is not the prefix itself that perfectivizes the simple verb *dat'* "give" without deriving a new lexeme.

I differentiated between the presence of a prefix and a perfectivizing prefix to control for potential similarities in the prefixational systems between German / Hungarian and Slavic due to lexical factors other than aspectuality. In all the four languages addressed, prefixation functions to derive (synchronically and diachronically) lexemes that are lexically more complex. This distinction was made to ensure that the distribution of prefixes in the four languages is not due to lexical effects other than aspectuality.

While that distinction is crucial with respect to the analysis of single verb forms, the two parameters did not influence with respect to the tests applied in this study. Therefore, the following sections will only list the parameter "presence of prefix".

Due to its lack in German and Hungarian[9], the presence of an imperfectivizing suffix has only been considered for Russian and Czech.

For coding purposes, I distinguish between four tenses: present, past, future, and infinitive[10], the latter referring to dependent infinitives. For German, I additionally distinguished between preterit and perfect, however, it did not show any effect and will not be considered in the remainder of this paper. For mood, indicative, imperative, and irrealis have been annotated.

Transitivity is tied to telicity, since direct objects often delimit the situation. Therefore, the presence (yes) or absence (no) of an accusative object was annotated to consider the transitivity of the verb (the same notation was applied to the presence/absence of a suffix, prefix, and the negation). Table 2 on page 191 summarizes the most important factors with their values.

The following section will elaborate on how the values for the factors aspect and actionality have been annotated in the four languages.

4.2 Annotation of aspect and actionality

Since aspect in Russian and Czech is systematically marked, its value could simply be determined by the form of the verb. In Hungarian and German, on the

9 As was mentioned in section 3.1, Hungarian has several verbal suffixes that change the actionality of the verb, e. g. derive frequentatives. As this is no systematic process (different suffixes, different compatibilities with verb roots, a high number of lexicalized forms), it has not been considered here.

10 The infinitive was grouped with other tense values for practical rather than linguistic reasons.

Table 2: The factors relevant to aspect marking.

factor	value
actionality	relative-stative (relstative) activity gradual-terminative (gradual) total-terminative (total)
aspect	perfective (pfv) imperfective (ipfv)
presence of a prefix	y n
presence of a suffix	y n
negation	y n
tense	present past future infinitive
mood	indicative imperative irrealis
presence of an accusative object	y n

other hand, aspectuality could not be expected to be marked in a systematic way. Therefore, the context of the situation was taken into account to determine whether a given token refers to a situation as temporarily bound (pfv) or unbound (ipfv).[11] Examples for this classification of predicates in German and Hungarian from the corpus are given in (10) and (11).

(10) Wir graben hier. (ipfv)
we dig.PRS.1PL here
'We mine here.'
(German, Noah)

(11) Senki sem fogja **meg**-látni. (pfv)
no.one NEG will.3SG.DEF **PFX**-see
'No one will see it.'
(Hungarian, Black Swan)

11 Inevitably, this choice is also influenced by e. g. the actiontality of the verb, tense and mood marking, and might vary across annotators. This issue cannot be taken up here, but should be addressed in a future study, e. g. in form of inter-rater agreement, in order to ensure the validity of such subtle semantic judgments.

We will now turn to a more detailed discussion of the four values of actionality adopted from Breu (1994, 2000). Since total-statives (e. g. *weigh*, *be called*) do not have a perfective counterpart in Slavic, only relative-statives have been considered in this study. Relative-stative predicates are defined by the following semantic properties: (i) the situation can but does not have to be inalienably bound to its participants; (ii) a temporal delimitation is possible, but not implied; and (iii) no supply of energy is required to maintain the situation.

Example (12) from the corpus illustrates a relative-stative verb.

(12) Either way, you'll shine. (Black Swan)

a. Tak ili inače, no ty budeš' blistat'.
like.this or like.that but you will.2SG shine.IPFV
(Russian)

b. Ať to dopadne jakkoli, budeš zářit.
whether it turn.out.PFV.3SG so will.2SG shine.IPFV
(Czech)

c. Így vagy úgy, de ragyogni fogsz.
like.this or like.that but shine. will.2SG
(Hungarian)

d. Auf die ein oder andere Weise, du
on the one or other way you
wirst auf der Bühne strahlen.
will.2SG on the stage shine
(German)

The next value of actionality, activity, corresponds to activities in the Vendlerian sense. It comprises situations that (i) are non-culminating, homogeneous.; (ii) with a possible but not implied temporal delimitation, and (iii) require a constant supply of energy to maintain the situation, as in (13) below:

(13) I was dancing the White Swan. (Black Swan)

a. Ja tancevala partiju beloj lebedi.
I dance.IPFV.PST part.ACC white.GEN swan.GEN
(Russian)

b. Tancovala jsem roli bílé labutě.
dance.PTCP was.1SG role.ACC white.GEN swan.GEN
(Czech)

c. Én táncoltam a Fehér Hattyút.
I dance.PST.1SG the white swan.ACC
(Hungarian)

d. Ich tanzte den weißen Schwan.
I dance.PST.1SG the.ACC white.ACC swan.ACC
(German)

We will now turn to total-terminatives. They are similar to what is traditionally understood as achievement verbs, although some differences exist. The semantic criteria for total-terminatives are: (i) the situation is culminating; (ii) a temporal delimitation is inherently given by the lexical semantics; (iii) the situation is not necessarily punctual. An example for the verb 'kill' is provided in (14).

(14) Are you here to kill me? (Inception)

a. Ty prišël ubit' menja?
you come.PFV.PST kill.PFV me.ACC
(Russian)

b. Jste zde, abyste mě zabil?
be.PRS.2SG here COMP.2SG me.ACC kill.PST.PTCP
(Czech)

c. Idejött, hogy meg-öljön?
here.come.PST.3SG COMP PFX-kill.COND.3SG
(Hungarian)

d. Sind Sie hier, um mich um-zu-bringen?
are you here for me.ACC PFX-to-kill
(German)

The last class of verbs considered are gradual-terminatives. Unlike the previous ones addressed, gradual-terminatives represent a complex class, i. e. consists of two phases. The first one is activity-like, but can lead to a point of culmination, the second phase, which is similar to total-terminatives. Aspectual marking can be used to point to either of the two phases: the imperfective highlights the activity (atelic) part, while the perfective aspect focuses on the culmination (the telic part). An example from the corpus is given below.

(15) open those gates (Frozen)

a.	otkroj	svoi	vorota
	open.**PFV**.IMP	your	gates.ACC
	(Russian)		
b.	otevřete	brány	
	open.**PFV**.IMP	gate.ACC	
	(Czech)		
c.	nyisd	**ki**	kapuidat
	open.IMP	**PFX:OUT**	gate.yours.ACC
	(Hungarian)		
d.	öffnet	die	Tore
	open.IMP	the.ACC	gates.ACC
	(German)		

Breu (1994, 2000) distinguishes another class of inchoative[12] verbs which consist of three phases. Since, even in Slavic, this class seems to comprise only a few lexemes due to its specific semantic requirements, this class of predicates will not be considered in the present study.

5 Results

This section discusses the findings of the corpus study. Section 5.1 addresses the distributions of the raw frequencies of imperfective and perfective forms in Russian and Czech, as well as the distribution of verbs with and without prefixes in Hungarian and German. Then, the importance of the factors to aspect marking will be addressed in section 5.2, as well as the similarity between the four languages with respect to aspect marking (section 5.3).

5.1 General distributions

In order to compare the marking of aspectuality between Slavic, Hungarian, and German, in this section, the occurrence of imperfectives and perfectives in Russian and Czech will be compared to the distribution of verbs without and verbs with prefixes in Hungarian and German. Figure 1 below shows the distribution for pfv/ipfv forms across the actional classes for Russian and Czech, as well as the presence (y) and absence (n) of a verbal prefix in Hungarian and German.

12 Note that the notion of 'inchoative' here is not used in the traditional way, for more details, see Breu (1994, 2000).

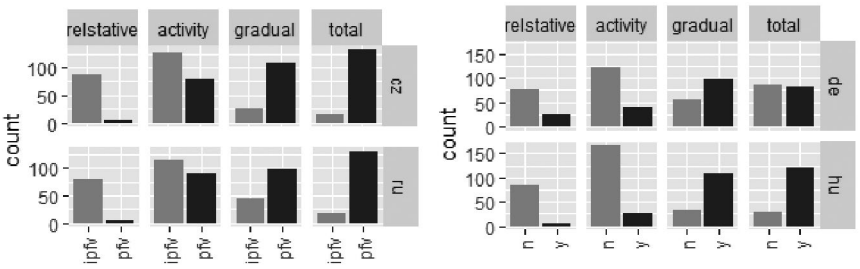


Figure 1: Aspectual values / presence of the prefix across actional classes.

In both Slavic languages, the perfective and imperfective forms occurred as expected: relative-stative verbs are almost exclusively imperfectives (the few perfective forms found were imperatives), also holding for activities as a weak tendency. Gradual-terminatives occurred more often as perfectives. As for total-terminatives, only a few instances of imperfectives are attested, almost all occurrences are perfectives. This reflects the compatibility of actionality and aspectual values: the two atelic classes (relative-stative, activity) are inherently more compatible with the imperfective value, hence, it is more frequent. The telic classes (gradual-terminative and total-terminative), on the other hand, are more compatible with the perfective value, the one attested in most instances.

As for Hungarian, the distribution of the prefix seems to follow the distribution of the aspectual forms in Slavic: almost no prefixes for relative-statives, and a strong trend for prefixed forms with gradual-terminatives and total-terminatives. Activity verbs seem to be less compatible with prefixes than imperfectives in Slavic, which also holds for German. Example (16) shows that the activity verb 'help' with a future meaning is perfective in Slavic, but lacks a prefix in Hungarian and German.

(16) Will He help us? (Noah)

a. On nam pomožet?
he us.DAT help.**PFV**.FUT.3SG
(Russian)

b. Pomůže nám?
help.**PFV**.3SG us.DAT
(Czech)

c. Segít rajtunk?
help.PRES.3SG us.SUPERESS
(Hungarian)

d. Wird Er uns helfen?
will.3SG he us.DAT help
(German)

Also in German, gradual-terminatives tend to be more frequent with prefixes than without. The following example shows how the situation is expressed by a perfective form in Slavic, and features a verbal prefix in Hungarian and German:

(17) You slipped on ice. (Frozen)

a. Vy poskol'znulis' na l'du.
you.POL slip.**PFV**.PST.REFL on ice.ACC
(Russian)
b. Uklouzl jste na ledu.
slip.**PFV**.PTCP AUX.2SG on ice.ACC
(Czech)
c. Csak **meg**-csúszott!
only **PFX**-slip.PST.3SG
(Hungarian)
d. Du bist **aus**-gerutscht.
you AUX.2SG **PFX**-slip.PTCP
(German)

As for total-terminatives, German contrasts with Hungarian and Slavic; both forms with and without prefixes occur with no preference. This suggests that there is no aspectual function involved in the prefixation for this class of verbs in German. In (18) below, Hungarian patterns with Slavic perfectives which have a prefix, while German has a simple verb.[13]

(18) How did you find it? (Inception)

a. Kak vy eë našli?
how you her.ACC find.**PFV**.PST
(Russian)
b. Jak jste to našel vy?
how AUX.2SG it.ACC find.**PFV**.PTCP you
(Czech)

13 The perfect marker *-ge* in German is not considered as a prefix that can be linked to aspect marking for the purposes of the present paper.

c. Hogyan találta **meg**?
how find.PST.3SG.DEF **PFX**
(Hungarian)

d. Wie haben Sie es gefunden?
how AUX.2SG you it.ACC find.PTCP
(German)

Looking at the distribution of perfective and imperfective forms across tenses and dependent infinitives in Figure 2, both Slavic languages prefer imperfectives in the present tense (perfective forms were only found in imperatives which were marked as *present* for tense); future forms and infinitives occurred almost only with perfectives, while the past tense showed a tendency for perfectives, also occurring with imperfectives.

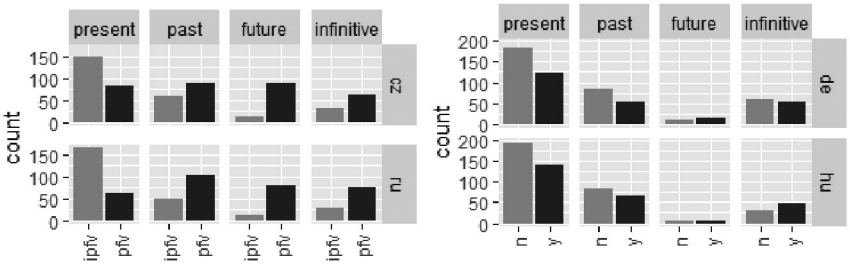


Figure 2: Aspectual marking / presence of the prefix across tense.

We find no strong trends for prefixation across tense in German or Hungarian. In general, prefixes are less available in the two languages irrespectively of aspectual functions. Only in Hungarian infinitives are verbs with prefixes more frequent than without. This can be accounted for by the function of dependent infinitives which often refer to situations as a whole, which in turn matches the limiting function of the prefix.

This contrast with Slavic is illustrated in example (19) below, showing present imperfectives of a gradual-terminative verb in both Russian and Czech, whereas Hungarian and German feature a prefix.

(19) Fire consumes all. (Noah)

a. Ogon’ vsë požiraet.
fire all consume.**IPFV**.PRS.3SG
(Russian)

b. Oheň vše ničí.
fire all destroy.**IPFV**.PRS.3SG
(Czech)

c. A tűz mindent **fel**-emészt.
the fire all **PFX:UP**-process
(Hungarian)

d. Feuer **ver**-zehrt alles.
fire **PFX**-consume all
(German)

Transitivity, annotated here as the presence (y) and absence (n) of an accusative object in a given instance, is expected to have an effect on prefixation in German and less so in Hungarian, whereas it should not play a role in aspect marking in Russian and Czech. Figure 3 shows the distribution of (im)perfective verbs and verbs with(out) prefixes across the presence of an accusative object.

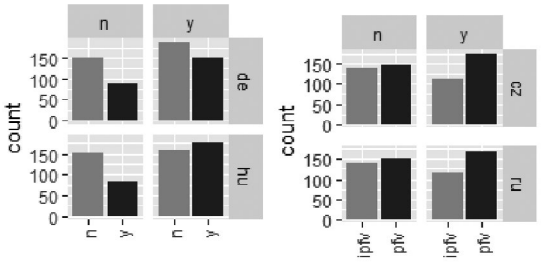


Figure 3: Perfectivity / presence of a prefix with transitivity.

As for verbs with an accusative object, no strong preference can be observed in Hungarian and German. Intransitive verbs, however, occurred with verbs without prefixes more frequently, which suggests that transitivity has the effect of making prefixes be more available to verbs.

Similarly, we do not find an effect for Slavic intransitives. Transitive verbs, on the other hand, show a very weak trend towards perfective forms.

5.2 Factor importance for the marking of aspectuality

After looking at the raw frequency distributions of aspectual forms in Slavic and the verbal prefix in Hungarian and German, we will now address the factors annotated and their importance with respect to the expression of aspect in the

Table 3: Performance of the random forest model for Russian, Czech, Hungarian, and German

Russian			**Czech**		
	Reference			Reference	
Prediction	ipfv	pfv	Prediction	ipfv	pfv
ipfv	213	12	ipfv	211	22
pfv	43	309	pfv	41	302
Accuracy: 0.9047			Accuracy: 0.9006		
No Information Rate: 0.5563			No Information Rate: 0.5625		
Hungarian			**German**		
	Reference			Reference	
Prediction	ipfv	pfv	Prediction	ipfv	pfv
ipfv	237	20	ipfv	218	37
pfv	24	291	pfv	23	299
Accuracy: 0.9231			Accuracy: 0.896		
No Information Rate: 0.5437			No Information Rate: 0.5823		

four languages. I used a random forest model to measure the importance of the factors.

Random forests (e. g. Baayen & Tagliamonte 2012; Baayen et al. 2008) can help to determine the strength of factors, i. e. to what extent they are correlated with the dependent variable (aspect). Random forests are based on a large number of conditional inference trees of random sub-samples of the data. Trees split the data according to the factor that makes the purest groups with the smallest p-value with respect to the dependent variable. Random forests (a large number of trees) have some advantages that are crucial for this study. They allow to control for factors that influence each other, as, e. g. tense and actionality, and to observe smaller effects, which would be hidden by more influential factors otherwise. As was noted in section 4.1, the factors considered for the models are: actionality, presence of a prefix, presence of a suffix, negation, tense, mood, and the presence of an accusative object.

Before turning to the results, the accuracy of the model will be addressed, i. e. the question of how well the model is able to capture the data. This is important, since it provides information on how reliable the results of the model are. To determine its accuracy, we let the model predict the values of the dependent variable (perfective, imperfective) based on the factors annotated. These predictions are then compared to the attested forms, providing information as to how well the model performs. Table 3 shows this by way of a confusion matrix for each forest modelling aspect marking in Russian, Czech, Hungarian, and German. The confusion matrix shows the number of tokens the model predicts as pfv/ipfv, while the reference marks the number of attested tokens. Taking Russian as

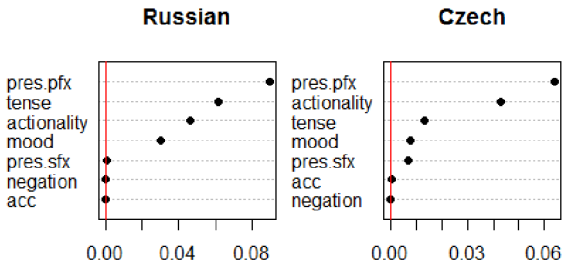


Figure 4: Conditional variable importance for aspect in Russian and Czech.

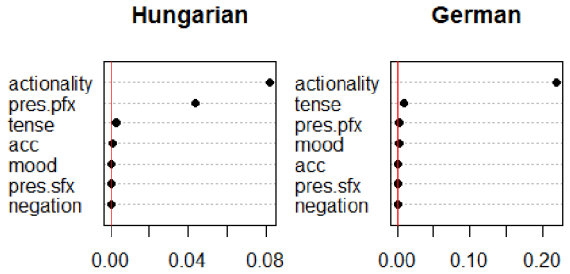


Figure 5: Conditional variable importance for aspect in Hungarian and German.

an example, the model correctly predicts 213 tokens as imperfectives, while 12 perfectives were predicted to be imperfectives. As for perfectives, the model correctly identified 309 tokens, but predicted 43 imperfective forms to be perfective. Given that the model is able to predict the majority of tokens correctly and the accuracy of 0.9047 being clearly above the no information rate[14], we can assume that the random forest for Russian with the factors considered is able to capture aspect marking. The same holds for the other languages, with an accuracy of approx. 0.9. This means that we can model the marking of aspectuality with the same factors in the four languages.

In Figures 4 and 5, we see the conditional variable importance of the factors examined. The conditional variable importance (e. g. Strobl et al. 2008, Baayen & Tagliamonte 2012) indicates how strongly a given factor is correlated with aspect. It is determined by randomly permuting the values of a single factor so that it is no longer linked to aspect. Then, the model's performance is tested: the greater the effect, i. e. the loss of accuracy, the higher the factor's importance.

14 The No Information Rate is the accuracy the model would have with the levels of the factors randomly manipulated.

The numbers in Figure 4 should not be understood in an absolute way, but are to be interpreted relative to each other. The red line marks significance.[15] The factors that fall to the left of it can be excluded from having a significant effect on the marking of aspect; the factors to the right show a significant correlation with the expression of aspect.

In both Russian and Czech, the presence of the prefix, followed by actionality, tense, and mood are relevant factors to predict whether an instance of a verb is likely to be perfective or imperfective. In Czech, in addition, the presence of a verbal suffix is significant. A more detailed discussion of the role of suffixes in Czech compared to Russian would surpass the scope of the present paper; however, it should be noted that previous work has argued for suffixation to be more productive in East-Slavic than West-Slavic (e. g. Wiemer & Seržant Forthc.; Arkadiev 2015). The present results rather suggest the opposite; as this surpasses the scope of the present paper, this issue will not be discussed in more detail here.

The presence of an accusative object and negation do not have an influence on the marking of aspect in either Russian or Czech. Hence, aspectual marking is highly correlated to the presence of a prefix, to lexical properties of the verb (actionality), and to other verbal categories.

In Hungarian and German, on the other hand, the most influential factor clearly is actionality. Hungarian shows a hybrid-like behaviour. On the one hand, it shares the high significance for actionality and significance of a much lower degree for tense with German. On the other hand, Hungarian patterns with Slavic for the high significance of the presence of the prefix to aspect marking. Thus, verbal prefixation in Hungarian is systematically involved in aspectual marking.

These findings support the initial hypothesis that aspect is systematically expressed in Hungarian to a certain extent, while it is not in German, so that Hungarian can be positioned between German (no aspect marking) and Slavic (aspect as a grammatical category). However, Hungarian also patterns with German in contrast to Russian and Czech, since the main factor correlated to aspectual functions is lexically determined. Although actionality plays a significant role in Slavic as well, it is less relevant than in Hungarian or German. Moreover, Slavic showed a significant influence for mood and tense, which means that aspect interacts with other verbal categories. For both German and Hungarian, the system is less complex as it depends on fewer factors and is more directly correlated to actionality, the lexical properties of the verb.

15 Following Strobl et al. (2008) to determine which factors are significant, their values were compared to the absolute value of the lowest negative value, the latter being indicated by the red line.

5.3 Similarity between the four languages with respect to aspect marking

Since the previous section showed that Hungarian patterns with German but also with Slavic with respect to different properties, this section will address the similarity between the four languages with respect to aspect marking in more detail. The similarity is determined also based on the factors considered for random forests, repeated here: aspect, negation, tense, mood, acc, presence of the prefix, presence of the suffix. Taking these factors in the four languages, we can measure the difference between them by clustering the languages according to their value distributions of the factors.

The cluster in Figure 6 confirms the results discussed in the previous sections. Russian and Czech pattern together, however, cutting the cluster at a higher point, Hungarian also patterns with Slavic, being situated between Slavic and German.

However, if we look at the four actional classes separately, we find that the languages cluster in two different ways. For activity and gradual-terminative verbs (see Figure 7), we find a cluster of Slavic on the one hand, and German and Hungarian on the other. Relative-stative and total-terminative verbs in Figure 8, however, show that Hungarian clearly patterns with Slavic instead of German.

We will now consider the clustering for each class in more detail. In section 5.1, it was observed for activity verbs that in Hungarian and German, many verbs do not combine with a prefix, so that there is no formal opposition available, which sets them apart from Slavic, featuring both imperfective and perfective forms (cf. (16)). This can explain why German and Hungarian pattern together for activity verbs.

Gradual-terminatives cluster in the same way, although they are compatible with both perfective and imperfective values and would be expected to be the first group of verbs showing aspect marking in an emerging aspect system, so that Hungarian would have been expected to pattern with Slavic for this group. A possible explanation for the clustering with German could be that, although in both languages prefixes are available, their distribution differs from perfective and imperfective forms in Slavic. Example (19) in section 5.1 showed that Slavic used the imperfective due to the present tense, whereas in both Hungarian and German the prefix was present. We also find cases in which Slavic uses a perfective form, with no prefix being present in Hungarian and German, possibly because of the direct object delimiting the situation, as in (20) below.

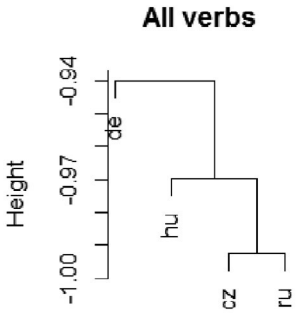


Figure 6: Similarity of the four languages for all verbs.

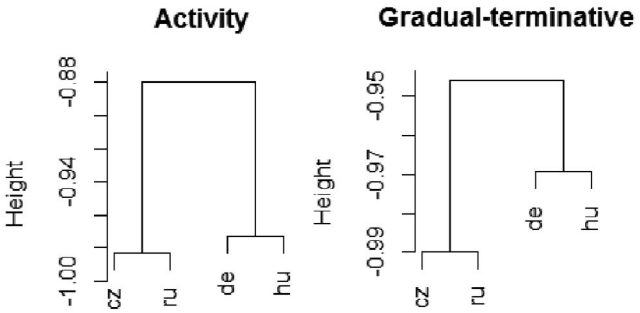


Figure 7: Similarity of the four languages for activity and gradual-terminative verbs.

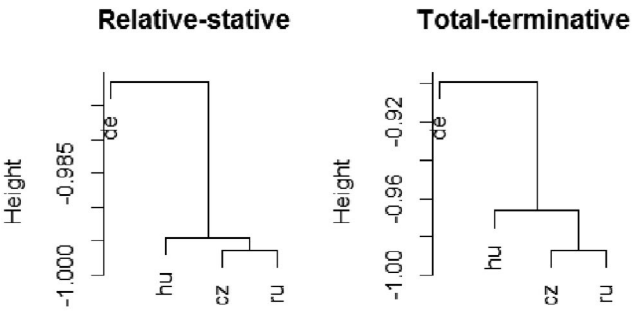


Figure 8: Similarity of the four languages for relative-stative and total-terminative verbs.

(20) Do you want to build a snowman? (Frozen)

a. ty xočeš' slepit' snegovika?
you want.PRS.2SG build.PFV snowman.ACC
(Russian)

b. Postavíme sněhuláka?
build.PFV.1PL snowman.ACC
(Czech)

c. Építünk hóembert?
build.PRS.1PL snowman.ACC
(Hungarian)

d. Bauen wir einen Schneemann?
build.PRS.1PL we a.ACC snowman.ACC
(German)

This means that although prefixation for gradual-statives is available in both German and Hungarian, the distribution of prefixes rather depends on actional properties and does not correspond to the aspectual distribution of Slavic forms.

Figure 8 illustrates that verbs from the relative-stative and total-terminative classes cluster Hungarian together with Slavic against German. For relative-statives, this can be explained by the fact that German uses verbal prefixes much more frequently for relative-statives than the other three languages (cf. Figure 3 in 5.1).

For total-stative verbs, section 5.1 (cf. Figure 1) showed that the distribution of the prefix in Hungarian follows the distribution of perfectives in Slavic, while there was no tendency for prefixation found in German, which can explain the cluster in Figure 8. An example where Hungarian patterns with Slavic featuring a prefix vs. German using a simple verb is given below:

(21) Although, I dreamed I was kissed by a troll. (Frozen)

a. […] što menja poceloval troll'.
COMP me.ACC kiss.**PFV**.PST troll
(Russian)

b. [...] že mě políbil troll.
COMP me.ACC kiss.**PFV**.PTCP troll
(Czech)

c. […] **meg**-csókolt egy troll.
PFX-kiss.PST.3SG a troll
(Hungarian)

d. [...] ein Troll hat mich geküsst.
a troll AUX.3SG me.ACC kiss.PTCP
(German)

6 Conclusion

This study addressed the systematicity of the expression of aspectuality in Hungarian compared to Russian, Czech, and German. The first two languages represented the East- and West-Slavic type of aspect respectively, while German functioned as the control language with verbal prefixation available, but without aspect marking.

Based on parallel subtitles, this study could empirically show to what extent Hungarian marks aspectuality by verbal prefixation and in which properties it resembles more Slavic or German behaviour. In addition, the verbs considered were split into four different actional classes to account for the interaction of inherent lexical (actional) properties and aspectual functions.

With respect to actional classes, the raw distributions of the prefix in German and Hungarian suggested that, except for activity verbs, prefixation in Hungarian indeed resembles the distribution of perfective and imperfective forms in Slavic. In German, on the other hand, especially for gradual-statives and total-statives, no such effect could be found. Distance-based similarity measures for the actional classes confirmed that Hungarian clusters with Slavic for relative-statives and total terminatives, while it showed that Hungarian forms a cluster with German not only for activities but also gradual-statives. Although prefixes are available in this class and it is semantically most compatible with both the perfective and imperfective values, the distribution of the prefix in both Hungarian and German is much more dependent on actionality and telicity, which is not the case in Slavic.

Based on random forest models, the importance of the factors annotated (actionality, presence of prefix, presence of suffix, tense, mood, presence of accusative object, negation) was determined. For all the four languages, the accuracy of the model was above 0.89, which means that the factors considered indeed capture the expression of aspect. With respect to the relevance of the factors, Hungarian showed a hybrid behaviour between German and Slavic, which could be confirmed by distance-based similarity measures for the four languages. What grouped it together with German was the fact the main significant factor to predict the aspectual value of a given form was actionality, i. e. an inherent lexical property, which argued against aspect as a grammatical category, systematically expressed and independent of lexical properties. However, actionality was significant in Slavic as well, amongst the presence of the prefix, tense, mood (and the presence of the suffix in Czech). This showed that in Slavic, aspect marking was sensitive to the actional properties of the verb as well. The presence of a prefix was also a highly significant factor in Hungarian, which made it group together with Slavic, whereas the presence of a verbal prefix in German, as expected, did not seem to correlate with aspectual values.

To conclude, this paper showed how the prefixation in Hungarian has to be situated between Slavic and German with respect to aspect marking. It could be shown that prefixation in Hungarian significantly correlates with the expression of aspectuality across the four actional classes. However, lexical properties of the verb, i. e. actionality, still have the greatest influence on the aspectual interpretation of a verb, which argues against a grammatical category of aspect in Hungarian.

References

Arkadiev, Peter. 2014. Towards an Areal Typology of Prefixal Perfectivization. *Scando-Slavica* 60 (2): 384–405.

Arkadiev, Peter. 2015. *Areal'naja tipologija prefiksal'nogo perfektiva (na materiale jazykov Evropy i Kavkaza).* Moskva: Jazyki slavjanskoj kul'tury.

Baayen, R. Harald; Douglas J. Davidson and Douglas M. Bates. 2008. Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items. *Journal of Memory and Language* 59 (4): 390–412.

Baayen, R. Harald and Sali A. Tagliamonte. 2012. Models, Forests and Trees of York English: Was/Were Variation as a Case Study for Statistical Practice. *Language Variation and Change* 24 (2): 135–178.

Breu, Walter. 1994. Interactions between Lexical, Temporal and Aspectual Meanings. *Studies in Language* 18 (1): 23–44.

Breu, Walter. 2000. Zur Position des Slavischen in einer Typologie des Verbalaspekts (Form, Funktion, Ebenenhierarchie und Lexikalische Interaktion). In Walter Breu (ed.), *Probleme der Interaktion von Lexik und Aspekt (ILA)*, 21–54. Tübingen: Max Niemeyer.

Csató, Éva Á. 1994. Tense and actionality in Hungarian. In Rolf Thieroff and Joachim Ballweg (eds.), *Tense systems in European Languages*, 231–246. Tübingen: Max Niemeyer.

Dahl, Östen. 1985. *Tense and Aspect Systems.* Oxford: Blackwell.

Dickey, Stephen M. 2000. *Parameters of Slavic Aspect: A Cognitive Approach.* Center for the Study of Language and Information: Stanford, CA.

Eördögh, Miklós. 1986. Aspect or aspectuality in Hungarian. In Wolfgang Heydrich and Janos S. Petöfi (eds.), *Aspekte der Konnexität und Kohärenz von Texten*, 115–27. Hamburg: Buske.

É. Kiss, Katalin. 2006. The Function and the Syntax of the Verbal Particle. In Katalin É. Kiss (ed.), *Event Structure and the Left Periphery. Studies on Hungarian,* Vol.68, 17–56. Budapest: Springer.

Isačenko, Alexandr V. 1960. Grammatičeskij stroj russkogo jazyka v sopostavlenii s slovackim. Čast'vtoraja: morfologija [Grammatical system in Russian as opposed to Slovak. Part 2: Morphology]. Bratislava: Izdatel'stvo akademii nauk.

Kiefer, Ferenc. 2006. *Aspektus és akcióminőség: különös tekintettel a magyar nyelvre*. Budapest: Akadémiai Kiadó.

Ladányi, Mária. 2015. Particle Verbs in Hungarian. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen and Franz Rainer (eds.), *Word-Formation. An International Handbook of the Languages of Europe*. Vol. 1, 660–672. Berlin/Boston: De Gruyter Mouton.

Lehmann, Volkmar. 1999. Der Aspekt. In Helmut Jachnow (ed.), *Handbuch der Sprachwissenschaftlichen Russistik und ihrer Grenzdisziplinen*, 214–242. Wiesbaden: Harrassowitz.

Lehmann, Volkmar. 2009. Formal-Funktionale Theorie Des Russischen Aspekts. http://subdomain.verb.slav-verb.org/Aspekttheorie.html (28 May 2017)

Levshina, Natalia. 2016. Verbs of Letting in Germanic and Romance: A Quantitative Investigation Based on a Parallel Corpus of Film Subtitles. *Languages in Contrast*. 16 (1): 84–117.

Lüdeling, Anke. 2001. *On Particle Verbs and Similar Constructions in German*. Dissertations in Linguistics. CSLI: Stanford, CA.

Piñón, Christopher J. 1995. Around the progressive in Hungarian. In István Kenesei (ed.), *Approaches to Hungarian. Volume Five. Levels and Structures*, 53–92. Szeged: JATE.

Sasse, Hans-Jürgen. 2001. Recent Activities in the Theory of Aspect: Accomplishments, Achievements, or Just Non-Progressive State? In *Arbeitspapiere des Instituts für Sprachwissenschaft der Universität zu Köln*. Vol. 40.

Smith, Carlota S. 1997. *The Parameter of Aspect*. Dordrecht, Boston, London: Kluwer.

Soltész, Katalin J. 1959. *Az Ősi Magyar Igekötők. Meg, El, Ki, Be, Le, Fel*. Budapest: Akadémiai Kiadó.

Stiebels, Barbara. 1996. *Lexikalische Argumente Und Adjunkte: Zum Semantischen Beitrag von Verbalen Präfixen Und Partikeln*. Berlin: Akademie Verlag.

Strobl, Carolin; Anne-Laure Boulesteix,, Thomas Kneib, Thomas Augustin and Achim Zeileis. 2008. Conditional variable importance for ran om forests. *BMC Bioinformatics* 9: 307.

Tatevosov, Sergej. 2002. The Parameter of Actionality. *Linguistic Typology* 6 (3): 317–401.

Wiemer, Björn. 2008. Zur Innerslavischen Variation bei der Aspektwahl und der Gewichtung ihrer Faktoren. In Karl Gutschmidt, U. Jekutsch, Sebastian Kempgen and Ludger Udolph (eds.), *Deutsche Beiträge Zum 14. Internationalen Slavistenkongreß, Ohrid 2008*, 383–409. München: Sagner.

Wiemer, Björn and Ilja A. Seržant. Forthcoming. Diachrony and Typology of Slavic Aspect: What does morphology tell us? In Andrej Malchukov and Walter Bisang (eds.), *Unity and diversity in grammaticalization scenarios*. Berlin: Language Science Press.

I. Corpus-Based Grammar Research
Current Trends and Issues

Daniela Elsner

Empirisch basierte Überlegungen zu Ableitungen mit *-weise/-erweise*

Abstract In our article we show how a quantitative and qualitative corpus analysis can be enriched by other empirical methods to gain a more comprehensive insight into a somewhat neglected topic, namely adverbial word-formation with the suffix *-(er)weise*. Pursuing Elsner's (2015) idea that *-(er)weise* should better be understood as two suffixes we bring forward syntactic arguments by showing that the suffixations differ with respect to their base positions in the German middle field. The results of a survey indicate that the interpretation of certain *-weise*-suffixations differs depending on the position of the lexeme in the sentence. Most strikingly formations with specific nominal bases can appear immediately before an indefinite noun where they only denote a large amount (*haufenweise Bücher* 'heaps of books') and can hardly be interpreted as adverbials anymore.

Keywords Adverb, Adverbial, Wortbildung, Grundpositionen von Adverbialen

1 Einführung

Bis auf wenige Ausnahmen (Heinle 2004, Ros 1992, Ronca 1975) ist die Wortbildung der Adverbien in der Literatur bisher kaum eingehend behandelt worden (vgl. auch Altmann/Kemmerling 2005: 153f.), was vor allem mit der diachronen Entwicklung einzelner Lexeme erklärt werden kann. So ist bei vielen Adverbien synchron kaum mehr zu erkennen, wie sie sich historisch entwickelt haben. Beispielsweise ist das Adverb *heute* auf die NP **hiu dauga* ‚(an) diesem Tage' zurückzuführen (Pittner et al. 2015: 11). Adverbiale Phrasen neigen häufig dazu, als zusammengerückte Adverbien reanalysiert zu werden (vgl. Waldenberger 2015), und in einigen Fällen führt dies zur Herausbildung von adverbialen Affixen, die mehr oder weniger produktiv zur Ableitung neuer Adverbien eingesetzt werden. Eines dieser adverbialen Affixe, das Suffix *-(er)weise*, soll hier näher beleuchtet werden. Zu Beginn werden die Ergebnisse einer Korpusanalyse aus Elsner (2015) referiert, die in einem ersten Schritt Aufschluss darüber geben, welche wortbildungstechnischen Möglichkeiten überhaupt bestehen und wie produktiv das Suffix ist. Anschließend stehen

zwei Fragen im Mittelpunkt dieses Beitrags: Zum einen diskutieren wir aufbauend auf der qualitativen Einteilung der Korpusdaten, in welche syntaktischen Adverbialklassen die jeweiligen Ableitungen mit *-(er)weise* einzuordnen sind. Dabei wird sich zeigen, dass neben semantischen und morphologischen (vgl. Elsner 2015) auch syntaktische Gründe dafür sprechen, dass es adäquater ist, von zwei Suffixen, *-erweise* und *-weise*, zu sprechen, da die Ableitungen je nach Basis und Suffix unterschiedliche Grundpositionen im Mittelfeld einnehmen. Zum anderen werden die Ergebnisse einer Pilotstudie präsentiert, in welcher der Frage nachgegangen wurde, inwiefern Interpretationsunterschiede (gleicher) Adverbien mit verschiedenen Grundpositionen korrelieren. Die Umfrage liefert erste empirische Argumente dafür, dass die Position von bestimmten nominalen *-weise*-Ableitungen vor artikellosen Objekten (z. B. *kiloweise Schminke*) eine Umgebung ist, in der die Adverbien desemantisiert sind und sich wie Determinative verhalten. Wir gehen abschließend kurz auf die Vor- und Nachteile einer Analyse dieser Ableitungen als Köpfe einer funktionalen Kategorie ein. Unser Beitrag zeigt, wie eine quantitative und qualitative Korpusauswertung durch theoretische Überlegungen sowie weitere empirische Methoden ergänzt werden kann.

2 Quantitative und qualitative Korpusauswertung[1]

Die Datenbasis bildet ein virtuelles Korpus mit Ausgaben des *Mannheimer Morgens* aus dem Jahr 2011, welches 21,73 Millionen Wortformen beinhaltet und Teil des Deutschen Referenzkorpus' ist. Mit Cosmas II wurde nach allen Lexemen, welche die Graphemfolge *weise* beinhalten, gesucht. Nach einer anschließenden Durchsicht, bei der unbrauchbare Belege wie z. B. *verweisen* manuell aussortiert wurden, konnten 10.588 Tokens für die weitergehende Analyse extrahiert werden. Die Auswertung der Korpusdaten gibt zunächst einen Überblick über die möglichen Wortstämme, mit denen sich die Einheit *-(er)weise* verbinden kann, und es zeigt sich eine deutliche Differenzierung von Wortbildungsbasis und Suffix: *-erweise* verbindet sich mit Adjektiven (2.739) und *-weise* mit Substantiven (7.839 Tokens). Zur Bestimmung der Produktivität des Wortbildungstyps nutzen wir Baayens „category-conditioned degree of productivity" (Baayen 2001: 157). Sowohl die Type-Token-Relation als auch der Produktivitätsindex P[2] deuten darauf hin, dass der Wortbildungstyp mit adjektivischen Basen und dem Suffix *-erweise* produktiver ist.

1 Die folgende Darstellung beruht auf Elsner (2015).

2 Der P-Wert gibt in diesem Fall an, wie gut ein Muster (hier ‚X-*erweise*' bzw. ‚X-*weise*') zur Bildung neuer Formen gebraucht werden kann. Er ist der Quotient aus der Anzahl der Hapaxlegomena und der Anzahl der Token aller Lexeme im Korpus, die nach diesem Muster gebildet sind (vgl. Baayen 2001).

Tabelle 1: Ergebnisse der Korpusanalyse zur Wortbildungseinheit *-(er)weise.*

Basis	Types	Tokens	TTR	Hapaxe	P
Adjektive	102	2.739	3,72%	39	0,0142
Substantive	117	7.839	1,49%	39	0,0050
Verben	1	10	–	0	–

Im Folgenden gehen wir kurz auf die Binnendifferenzierung der substantivischen und adjektivischen Ableitungen ein. Substantivische *-weise*-Ableitungen können dahingehend unterschieden werden, ob sie über eine deverbale oder nicht-deverbale substantivische Basis verfügen. In Abhängigkeit von der Basis zeigen sich semantische Unterschiede:

(1) **Häppchenweise** gibt der Radiosender seine Preisträger bekannt: […]. (M11/FEB.03486)
(2) Der Cellist Tomasz Daroch, der in Mannheim studierte und gerade **vertretungsweise** zum Philharmonischen Orchester Heidelberg gehört, hat […]. (M11/FEB.04957)

Ableitungen mit deverbalen Substantiven wie in (2) denotieren typischerweise einen (abstrakten) Vergleich (‚als/zur Vertretung'). Demgegenüber sorgen Ableitungen mit nicht-deverbalen Substantiven typischerweise für eine Vereinzelung des vom Verb denotierten Prozesses.[3] Dies hat eine Quantifizierung zur Folge: Anstelle einer Bekanntgabe erfolgt in (1) eine Aufspaltung dieses Prozesses in einzelne (kleinere) Bekanntgaben. Gerade solche substantivischen Ableitungen finden sich häufig in der Position vor indefiniten, artikellosen Objekten und fungieren als Quantoren *(kistenweise Äpfel).* Andererseits zeigen sie sich seltener in einer attributiven Position zwischen Artikel und Substantiv, als es Ableitungen mit deverbalen Substantiven als Basis tun.

Neben der Tatsache, dass sich die Suffixe *-erweise* und *-weise* hinsichtlich ihrer Produktivität sowie ihrer Basis unterscheiden, spricht auch das obligatorische Auftreten des (ehemaligen) Flexionssuffixes *-er* dagegen, *-erweise* als eine verfugte Variante zu *-weise* zu interpretieren. Fuhrhop (1996: 525) weist darauf hin, dass Fugenelemente „keine[r] klare[n] Systematik" unterliegen; daher

3 Das zeigen die verschiedenen Type-Token-Relationen (vgl. Elsner 2015: 110). Die Kombination von deverbalem Basissubstantiv und vergleichender Bedeutung hat eine höhere Type-Token-Relation, nämlich 2,57 %, als die Kombination von deverbalem Basissubstantiv und vereinzelnder Bedeutung (0,46 %). Demgegenüber weist die Kombination von nicht-deverbalem Basissubstantiv und vergleichender Bedeutung eine Type-Token-Relation von 0,11 % auf, die Kombination von nicht-deverbalem Basissubstantiv und vereinzelnder Bedeutung 16,27 %.

wäre ein solches regelhaftes Auftreten von *-er* eher ungewöhnlich. Adjektivische *-erweise*-Ableitungen bilden typischerweise Satzadverbien; hierbei handelt es sich zwar um einen produktiven, aber nicht völlig unrestringierten Prozess, wie ungrammatische Bildungen wie **hoherweise* oder **lauterweise* zeigen. Unter anderem solche Adjektive, die eine rein modale Lesart haben, eignen sich offensichtlich nicht als Ableitungsbasis. Das Suffix *-erweise* kann sich jedoch auch mit (adjektivischen) Partizip I-Formen verbinden, und die Produkte dieser Kombination können nicht als Satzadverbien kategorisiert werden. Lexeme wie *lesenderweise* oder *tanzenderweise* nehmen Bezug auf einen Umstand, der den durch den Verbalkomplex denotierten Prozess begleitet. Im Folgenden zeigen wir, dass auch (weitere) syntaktische Gründe für die Differenzierung der beiden Suffixe sprechen: Je nach Basis haben die Lexeme verschiedene syntaktische Grundpositionen und sind somit verschiedenen Adverbialklassen zuzuordnen.

3 Positionen und Interpretationen

Nach ihrer Grundposition im Mittelfeld fallen Adverbiale in verschiedene Klassen (vgl. Frey/Pittner 1998, Frey 2003, Pittner 2004). Zu unterscheiden sind (in ansteigender Einbettungstiefe): Frame-/Bereichsadverbiale – Satzadverbiale – ereignisbezogene Adverbiale – ereignisinterne Adverbiale – prozessbezogene Adverbiale. Dabei ist strittig, ob die Grundposition von Satzadverbialen unter- oder oberhalb von Frame- und Bereichsadverbialen liegt (s. dazu auch Störzer/Stolterfoht 2013), ob Bereichsadverbiale zu den Frame- oder den Satzadverbialen gehören oder eine eigene Klasse bilden und ob prozessbezogene Adverbiale ihre Grundposition unter- oder oberhalb von Objekten haben (s. dazu z. B. Schäfer 2013).

Von den von Frey/Pittner (1998) zur Ermittlung der Grundpositionen vorgeschlagenen Tests können weder der Quantorenskopus noch die Prinzip-C-Effekte herangezogen werden.[4] Die Ergebnisse der verbleibenden Proben (Fokusprojektion, Thema-Rhema-Bedingung, komplexes Vorfeld und Stellungsfestigkeit existenziell interpretierter *w*-Phrasen) deuten darauf hin, dass adjektivische *-erweise*-Ableitungen als Satzadverbiale ihre Grundposition oberhalb der Subjekte haben, während partizipiale *-erweise*-Ableitungen als ereignisinterne Adverbiale zwischen Subjekt und Objekt basisgeneriert sind und substantivische *-weise*-Ableitungen als prozessbezogene Adverbiale am tiefsten eingebettet sind und ihre Grundposition unterhalb von Objekten haben.

4 Einerseits können *-(er)weise*-Ableitungen nicht mit Quantoren kombiniert werden, andererseits stellen sie keine R-Ausdrücke dar und können mit solchen daher nicht koindiziert sein.

Adj+*erweise* > Subjekt > PartI+*erweise* > Objekt > N+*weise*

(3) Möglicherweise hat Hans musikhörenderweise den Sand eimerweise gesiebt.

Für die substantivischen und partizipialen Ableitungen zeigt dies in (4) und (5) beispielhaft der Vorfeldtest, der darauf beruht, dass sich im Vorfeld keine ungebundenen Spuren befinden dürfen.

(4) a. [Häppchenweise gegessen]$_i$ hat Hans den Schweinebraten t_i.
b. ? [Den Schweinebraten t_j gegessen]$_i$ hat Hans [häppchenweise]$_j$ t_i.
(5) a. [Das Ufer erreicht]$_i$ hat Hans schwimmenderweise t_i.
b. ? [Schwimmenderweise t_j erreicht]$_i$ hat Hans [das Ufer]$_j$ t_i.

Als ereignisinterne Adverbiale können die partizipialen *-erweise*-Ableitungen verschiedene semantische Ausrichtungen haben. So wird in (6) ein Begleitumstand benannt: Der Mann fiel demnach auf, während er ein Auto (o. Ä.) fuhr. In (7) hingegen ist eine Interpretation als Instrument adäquater; der Ballon wirbt für die Stadt, indem er fährt.

(6) Gegen 20.25 Uhr fiel der Mann erneut **fahrenderweise** einer Streifenwagenbesatzung […] auf. (RHZ10/FEB.10420)
(7) Innsbruck hat einen neuen Ballon, der **fahrenderweise** für die Stadt werben soll. (I99/MAI.19381)

Eine formlose Befragung von Studierenden hat ergeben, dass das Suffix in diesen Fällen als überflüssig und die Sätze durchweg als markiert empfunden werden. Auch Ronca (1975) argumentiert für einen pleonastischen Status des Suffixes. An dieser Stelle sei zumindest darauf hingewiesen, dass das Suffix für einen eindeutigen Wortartenwechsel (Adjektiv > Adverb) sorgt und damit eine prädikative Lesart des Wortbildungsprodukts unterbindet[5], sodass nicht von einer völligen Funktionslosigkeit ausgegangen werden kann (vgl. auch Elsner 2015). Wir können bis hierher festhalten, dass die *-(er)weise*-Ableitungen in Abhängigkeit von ihrer Basis jeweils verschiedene Grundpositionen einnehmen. Ableitungen mit dem Suffix *-weise* unterscheiden sich also syntaktisch von Ableitungen mit dem Suffix *-erweise*. Neben den in Elsner (2015) genannten morphologischen und semantischen Eigenschaften der jeweiligen Wortbildungsprodukte weisen demnach auch syntaktische Eigenheiten darauf hin, dass es adäquater ist, von zwei Suffixen zu sprechen und nicht von einem Suffix, das in bestimmten Kontexten eine verfugte Variante hat.

5 Ähnlich argumentiert Van de Velde (2005) für das niederländische Pendant *-erwijs*.

Im Folgenden werden die Ergebnisse einer Pilotstudie vorgestellt, bei der mithilfe einer Umfrage überprüft wurde, inwiefern Interpretationsunterschiede gleicher Adverbien mit verschiedenen Positionen korrelieren. In Elsner (2015) wird die These aufgestellt, dass die Interpretation spezifischer substantivischer *-weise*-Ableitungen[6] abhängig ist von ihrer Position. Es gilt:

(I) [Container/Maß+*weise*] vor nicht deverbalen Nomina denotieren eine große Menge (*kistenweise Wein* = ‚viel Wein')
(II) [Container/Maß+*weise*] vor deverbalen Nomina spezifizieren den Prozess (*die kistenweise Lagerung* = ‚die Lagerung in Kisten')
(III) [Container/Maß+*weise*] vor einem Verb spezifizieren den Prozess (*die Bananen kistenweise lagern* = ‚Bananen in Kisten lagern')

Diese Thesen wurden mithilfe einer Umfrage überprüft, bei der 76 Probanden verschiedene (konstruierte) Sätze erhielten, die im Hinblick auf ihre Grammatikalität bewertet werden sollten. Dabei steht der Wert 1 für sehr schlecht und der Wert 5 für sehr gut. Zusätzlich sollten Fragen zur Bedeutung der Sätze beantwortet werden. So musste beispielsweise für (8a, b) jeweils angekreuzt werden, ob (i) der Wein sich in Flaschen befindet (= Containerlesart), (ii) es sich um eine große Menge Wein handelt (= quantifizierende Lesart), (iii) beides (also (i) und (ii)) oder (iv) nichts von all dem zutrifft.

(8) a. Er hat eine Alkoholvergiftung erlitten, weil er flaschenweise Wein getrunken hat.
b. Ich habe gehört, dass man im Großhandel Wein flaschenweise kaufen kann.

Tabelle 2: Ergebnisse der Umfrage zu Interpretationsunterschieden.

	[Cont./Maß+weise] + nicht deverbales Nomen	[Cont./Maß+weise] + deverbales Nomen	[Cont./Maß+weise] + Verb
quantifizierende Lesart	51 (67 %)	40 (53 %)	15 (20 %)
Containerlesart	–	36 (47 %)	19 (25 %)
beide Lesarten	25 (33 %)	–	41 (54 %)
keine der Lesarten	–	–	1 (1 %)

Bezüglich der Akzeptabilität zeigt sich, dass Konstruktionen mit [Container/Maß+*weise*] vor deverbalen Nomina *(das flaschenweise Trinken, die säckeweise Lagerung)* als weniger gut (3,3) empfunden werden als die beiden anderen

6 Darunter fallen solche Ableitungen, deren substantivische Basis einen Container oder ein Maß bezeichnen.

Konstruktionen (beide 4). Was die Bedeutung betrifft, geben 67 % der Probanden an, dass die Ableitungen in der Konstruktion [Container/Maß+*weise*] vor nicht deverbalen Nomina eine rein quantifizierende Lesart haben (*kistenweise Wein* = ‚viel Wein'), für 33 % liegt eine Kombination aus quantifizierender und Containerlesart vor (*kistenweise Wein* = ‚viel Wein, der in Kisten gelagert wird'). Bei der Konstruktion [Container/Maß+*weise*] vor deverbalen Nomina gibt es kein eindeutiges Ergebnis. Die rein quantifizierende und die reine Containerlesart werden zu je ca. 50 % als präferierte Lesart angegeben. Hinsichtlich der dritten Konstruktion, [Container/Maß+*weise*] vor Verben, ist zumindest die rein quantifizierende Lesart selten (20 % der Probanden entschieden sich dafür). Die meisten gaben an, dass die Ableitungen hier sowohl eine quantifizierende als auch eine Containerlesart haben (*Bananen kistenweise lagern* = ‚es werden viele Bananen in Kisten gelagert'). Die in Elsner (2015) aufgestellten Thesen können damit zunächst nur teilweise empirisch verifiziert werden. Recht eindeutig zeigt sich jedoch, dass die Ableitungen vor nicht deverbalen Nomina eine quantifizierende Bedeutung tragen und damit desemantisiert sind. In dieser Position lassen sie sich zudem keiner der gängigen Adverbialklassen zuordnen, und es stellt sich die Frage, wie sie adäquat analysiert werden können. Dass diese Ableitungen auch in eingebetteten NPs (9) auftreten können, verdeutlicht, dass zum Verb kein struktureller Bezug mehr vorhanden ist.

(9) Mit **tonnenweise** Schlafsäcken und Matratzen wollen mehrere europäische Staaten die Erdbebenopfer in Japan unterstützen. (M11/MAR.07998)

In dieser Position sind die Ableitungen nicht flektierbar, etwaige Adjektive müssen stark flektieren *(tonnenweise warme Schlafsäcke)* und das Auftreten von Artikeln ist blockiert (vgl. Elsner 2015). Auffällig ist, dass als Bezugssubstantive nur pluralische oder Massennomina möglich sind. Aufgrund dieser Merkmale liegt eine Interpretation der Ableitungen als Determinative (D^0) nahe – prinzipiell könnten die Ableitungen jedoch auch als Köpfe von Adjektivphrasen (AP), Quantifiziererphrasen (QP) oder Gradphrasen (DegP) aufgefasst werden. Wir werden diese Möglichkeiten kurz diskutieren.

Löbel (1990) führt die funktionale Kategorie Q ein, welche die Aufgabe hat, Nomina zählbar zu machen. Q^0-Elemente sind genau diejenigen Einheiten, die auch als Basis der hier zu Diskussion stehenden *-weise*-Ableitungen fungieren können (z. B. *drei **Kisten** Bier – kistenweise*). Zudem können Q^0-Elemente ebenfalls ausschließlich mit Massennomina und Nomina im Plural kombiniert werden. Gegen die Interpretation als Q^0 spricht jedoch die Tatsache, dass die *-weise*-Ableitungen gerade nicht dafür sorgen, dass Nomina zählbar gemacht werden (vgl. 10).

(10) *Hans möchte fünf kistenweise Wasser kaufen.

Darüber hinaus unterbinden Q^0-Elemente nicht das Auftreten eines Determinierers. Nach Bhatt (1990: 68) nimmt Deg^0 ausschließlich APs als Komplemente; die *-weise*-Ableitungen treten jedoch auch ohne APs auf und stehen nicht als Modifikatoren vor einem Adjektiv. Gegen die Interpretation als A^0 spricht, dass die Ableitungen nicht flektieren und es nur wenige Adjektive gibt, die dies nicht tun. Zudem sind die Ableitungen nicht wie Adjektive iterierbar *(*literweise flaschenweise Wein)* und sie können Adjektiven nur folgen, aber nicht vorangehen (**kaltes literweise(s) Wasser*). Insgesamt scheint die Interpretation als D^0 damit die adäquateste Lösung zu sein. Zwar trägt D^0 die Agreement-Merkmale, jedoch können Kasus und Numerus lexikalisch am Substantiv realisiert sein; Person und Genus sind den Substantiven inhärent, sodass lediglich das Merkmal Definitheit durch die *-weise*-Ableitungen realisiert wird, und zwar als [-def]. Da es sich bei den Bezugssubstantiven aber per se um indefinite Nomina handelt, sind sie streng genommen syntaktisch überflüssig. Es ist zu überlegen, ob ihnen eine rein semantische Funktion zukommt, die evtl. in einer starken Betonung der Vielheit liegt. Problematisch an der hier vorgeschlagenen Analyse ist sicherlich, dass D^0 als funktionale Kategorie eine geschlossene Klasse darstellt. Als Alternative bleibt zu prüfen, ob die Ableitungen als Modifikatoren der NP aufzufassen sind, was auch ihre topologische Variabilität in Kontexten, in denen die NP nicht in eine PP eingebettet ist, erklären könnte. Topologische Eigenheiten weisen darauf hin, dass die *-weise*-Ableitung und das Bezugsnomen auch bei diskontinuierlicher Positionierung nicht als zwei unabhängige Einheiten aufgefasst werden können:

(11) a. Vor den Barrikaden haben sie eimerweise Wasser verschüttet.
b. Eimerweise Wasser haben sie vor den Barrikaden verschüttet.

(12) a. Hans hat flugs einen Blumenstrauß gekauft.
b. *Flugs einen Blumenstrauß hat Hans gekauft.

(12b) zeigt, dass Adverb und Substantiv normalerweise nicht gemeinsam im Vorfeld stehen können – die Kombination von *-weise*-Ableitung und Substantiv ist jedoch problemlos möglich (vgl. 11b). Eine Analyse der *-weise*-Ableitung als Modifikator der NP kann eventuell das Bewegungsverhalten besser erklären, muss aber auch eine Lösung dafür finden, dass D^0 unbesetzt bleiben muss.

4 Fazit

Die Suffixe *-weise/-erweise* gehören zu den produktivsten Suffixen, um Adverbien abzuleiten (vgl. Altmann/Kemmerling 2005: 167, Fleischer/Barz 2012: 369). Ausgehend von einer Korpusanalyse wurde gezeigt, dass sie sich in ihrer

Produktivität und die Wortbildungsprodukte sich hinsichtlich ihrer Ableitungsbasen, ihrer Semantik und ihrer Syntax unterscheiden. Je nach Art der Basis nehmen die Lexeme unterschiedliche adverbiale Grundpositionen ein: N+*weise*-Ableitungen sind als prozessbezogene, PartI+*erweise*-Ableitungen als ereignisinterne und Adj+*erweise*-Ableitungen als Satzadverbiale aufzufassen. Dies ist ein weiteres Argument für die These, dass es sich bei *-weise* und *-erweise* um verschiedene Suffixe handelt. Die Ergebnisse einer empirischen Pilotstudie zeigen, dass bestimmte substantivische Ableitungen in der Position vor Objekten primär eine quantitative Lesart aufweisen. Sie sind desemantisiert und können keiner der gängigen Adverbialklassen zugeordnet werden; vielmehr wurde dafür argumentiert, dass es sich um Determinative handelt.

Literaturverzeichnis

Altmann, Hans/Kemmerling, Silke (2005): Wortbildung fürs Examen. 2., überarbeitete Auflage. Göttingen: Vandenhoeck & Ruprecht.

Baayen, Harald (2001): Word Frequency Distributions. Dordrecht: Kluwer.

Bhatt, Christa (1990): Die syntaktische Struktur der Nominalphrase im Deutschen. Tübingen: Narr.

Elsner, Daniela (2015): Adverbial morphology in German: Formations with –weise/-erweise. In: Pittner, Karin/Elsner, Daniela/Barteld, Fabian (Hg.): Adverbs. Functional and diachronic aspects. (= Studies in Language Companion Series 170). Amsterdam: John Benjamins, S. 101–132.

Fleischer, Wolfgang/Barz, Irmhild (2012): Wortbildung der deutschen Gegenwartssprache. 4., völlig neu bearbeitete Auflage. Berlin: de Gruyter.

Frey, Werner (2003): Syntactic conditions on adjunct classes. In: Lang, Ewald/Maienborn, Claudia/Fabricius-Hansen, Cathrine (Hg.): Modifying Adjuncts. Berlin: de Gruyter, S. 163–209.

Frey, Werner/Pittner, Karin (1998): Zur Positionierung der Adverbiale im deutschen Mittelfeld. In: Linguistische Berichte 176, S. 489–534.

Fuhrhop, Nanna (1996): Fugenelemente. In: Lang, Ewald/Zifonun, Gisela (Hg.): Deutsch typologisch. Berlin: de Gruyter, S. 525–550.

Heinle, Eva-Maria (2004): Diachronische Wortbildung unter syntaktischem Aspekt. Das Adverb. Heidelberg: Winter.

Löbel, Elisabeth (1990): D und Q als funktionale Kategorien in der Nominalphrase. In: Linguistische Berichte 127, S. 232–264.

Pittner, Karin (2004): Where syntax and semantics meet: Adverbial positions in the German middle field. In: Austin, Jennifer R./Engelberg, Stefan/Rauh, Gisa (Hg.): Adverbials. The interplay between meaning, context, and syntactic structure (= Linguistik Aktuell/Linguistics Today 70). Amsterdam: John Benjamins, S. 253–287.

Pittner, Karin/Elsner, Daniela/Barteld, Fabian (2015): Introduction. In: dies. (Hg.): Adverbs. Functional and diachronic aspects. (= Studies in Language Companion Series 170). Amsterdam: John Benjamins, S. 1–17.

Ronca, Dorina (1975): Morphologie und Semantik deutscher Adverbialbildungen. Eine Untersuchung zur Wortbildung der Gegenwartssprache. Dissertation, Rheinische Friedrich-Wilhelms-Universität zu Bonn.

Ros, Gisela (1992): Suffixale Wortbildungsmorpheme. Untersuchungen zu ihrer semantischen Leistung am Beiwort der deutschen Gegenwartssprache. Stuttgart: Hans-Dieter Heinz Akademischer Verlag.

Schäfer, Martin (2013): Positions and Interpretations. German Adverbial Adjectives at the Syntax-Semantics Interface (= Trends in Linguistics. Studies and Monographs 245). Berlin: de Gruyter Mouton.

Störzer, Melanie/Stolterfoth, Britta (2013): Syntactic base positions for adjuncts? Psycholinguistic studies on frame and sentence adverbials. In: Questions and Answers in Linguistics 1(2), S. 57–72.

Van de Velde, Freek (2005): Exaptatie en subjectificatie in de Nederlandse adverbiale morfologie. In: Handelingen der Koninklijke Zuid-Nederlandse Maatschappij voor Taal- en Letterkunde en Geschiedenis 58, S. 105–124.

Waldenberger, Sandra (2015): Lexicalization of PPs to adverbs in historic varieties of German. In: Pittner, Karin/Elsner, Daniela/Barteld, Fabian (Hg.): Adverbs. Functional and diachronic aspects. (= Studies in Language Companion Series 170). Amsterdam: John Benjamins, S. 179–205.

Lea M. Fricke, Swantje Tönnis

Es ist dies – A Special Use of German Prefield-*es*[1]

Abstract We present a corpus study on a hitherto unstudied use of the German prefield-*es* in combination with a demonstrative subject *dies* and a copula verb *ist*, which we call *Es ist dies*-sentences. In such constructions, the prefield-*es* appears redundant as they contain a suitable and mostly preferred candidate to fill the prefield, the demonstrative pronoun *dies*. According to our corpus data, this construction is predominantly used in southern varieties of German (Swiss, Austrian and Bavarian German). In order to better understand the distribution of these constructions, we compared *Es ist dies*-sentences to a sample of unmarked *Dies ist*-sentences that mirrored the distribution of the prefield-*es* cases. We found two significant differences between the two samples with regard to a) the distance to the antecedent of *dies* and b) the content of the sentence. Based on our findings, we propose a modification of Speyer's (2008, 2009) stochastic Optimality Theoretic (OT) model of prefield ranking.

Keywords Prefield-*es*, information structure, stochastic OT, southern varieties of German

1 Introduction

The use of prefield-*es* in sentences like (1) seems redundant and therefore marked, as the sentence ostensibly offers a better candidate to fill the prefield position: the demonstrative pronoun *dies* ('this'). Thus, the version presented in (2) appears more natural.

1 Many thanks go to Edgar Onea for his valuable suggestions as well as his comments on earlier versions of this paper, and to Alexander Schreiber for advising us on the statistical analysis. We also want to thank the organizers of *Grammar and Corpora 2016* for providing the opportunity for such an inspiring exchange. We are grateful for the helpful comments we received from the audience at the poster session, we thank in particular Erik Fuß, Carlo Geraci, Marek Konopka, and Helmut Weiß. Finally, we thank two anonymous reviewers for their comments.

(1) **Es ist dies** der schwerste Fall von Marktmanipulation, den wir
It is this the most severe case of market manipulation that we
je gesehen haben.[2]
ever seen have
'This is the worst case of market manipulation that we have ever seen.'

(2) **Dies ist** der schwerste Fall von Marktmanipulation, den wir
this is the most severe case of market manipulation that we
je gesehen haben.
ever seen have
'This is the most severe case of market manipulation that we have ever seen.'

This construction is not an idiom, since it allows a significant range of variation. There is also a variant featuring *das*, 'that', instead of *dies*. Moreover, the construction may surface with different inflected forms, with and without a relative clause, and the NP can be preceded by further elements. In this paper, we focus on the variant presented in (1). After explaining in more detail why this construction violates expectations about the use of prefield-*es*, we present our corpus study, which is an investigation into the conditions of its use. Based on the results we present our tentative analysis, a modification of Speyer's (2008, 2009) prefield ranking.

2 Background

In a standard German declarative matrix clause, the finite verb occurs in the second position. This means that the prefield, the position in front of the finite verb, needs to be filled by one constituent. In some cases, exemplified by sentence (3a.), this is brought about by the non-phoric use of the third person neuter pronoun *es,* which does not contribute to the truth-conditions of the sentence. Unlike the also non-phoric subject-*es* which functions as a formal subject for verbs that do not assign thematic roles (e.g. *es regnet*, 'it is raining'), this so-called prefield-*es* is not an argument (see Pütz 1986, Tomaselli 1986, Cardinaletti 1990, Zifonun 1995, Paranhos Zitterbart 2002, and Pittner & Bermann 2004 for the different uses of *es*). It only serves to fill the prefield in order to have a verb second clause. The prefield is the only position this type of *es* can occur in, as shown by the ungrammatical example (3b.).

2 http://www.tagesanzeiger.ch/wirtschaft/unternehmen-und-konjunktur/Es-ist-der-schwerste-Fall-den-wir-je-gesehen-haben/story/31098247.

(3) a. Es kommen viele internationale Gäste.
it come many international guests
'Many international guests are coming.'
b. *Hoffentlich kommen es viele internationale Gäste.
hopefully come it many international guests
'Hopefully, many international guests are coming.'

The reason for using non-phoric *es* in the prefield instead of the subject has been argued to lie in the information structure. It has been suggested that *es* may be located in the prefield as a placeholder when the subject carries the informational load (Zifonun et al. 1997) and represents new information (Pittner & Bermann 2004) and therefore tends to be located towards the end of the sentence. Speyer (2009) further investigates the conditions that allow the occurrence of *es* in the prefield. He characterizes the use of prefield-*es* as a "last resort" to fill the prefield in order to have a V2-sentence if there is no better candidate available. A better candidate according to his stochastic OT based prefield ranking would be the topic of the sentence which Speyer (2009: 339) defines in terms of Centering Theory (Grosz et al. 1995; Walker et al. 1998) as a 'macrostructurally relevant' entity. This means a topic either needs to be discourse-old (i.e. it occurs in the directly preceding sentence, Speyer 2009: 336), or relevant in the further course of the text in order to be allowed in the prefield. In addition to lacking a more appropriate prefield filler, sentences featuring prefield-*es* were observed to contain few constituents, often only the subject (Speyer 2009: 334).

Clearly, in the last respect our *Es ist dies*-sentences differ from the classic cases of prefield-*es*, as they are copular sentences, which always contain at least two arguments. Furthermore, the fact that the phoric *dies*[3] refers back to an antecedent in the text indicates that it does not represent new information and seems to point to its macrostructural relevance. However, in a first explorative examination of *Es ist dies*-sentences in context, we observed many cases in which the antecedent is not located in the preceding sentence but found at a greater distance. The text passage in (4) exemplifies this.

(4) (Last Sunday, his majesty, Jens-Peter I, the current champion marksman, planted **his royal tree** in the castle garden in Warberg, accompanied by the former majesties of the Warberg shooting association. In the context of the re-development of the castle garden, the shooters had decided that every reigning majesty should plant his or her own tree.)

3 It is a characteristic of German demonstratives in copular sentences that they can remain uninflected. Diessel (1999) uses the label demonstrative identifier for this use of demonstrative pronouns.

Es ist dies nun der 17. Baum der seinen Platz im Park findet.[4]
it is this now the 17th tree that its place in the park finds
'This is now the 17th tree that finds its place in the park.'

Here, *dies* refers back to the tree that is mentioned in the beginning of the short passage, not to the one directly preceding the *Es ist dies*-sentence, as *his or her own tree* is within the scope of a quantifier and therefore cannot be the antecedent of *dies*. Mentioning that the tree just planted is the 17th tree in (4) is an instance of taking stock. Such sentences of the form *It is this the nth…* were found quite often. We also found a number of occurrences of *Es ist dies*-sentences that express evaluative comments, like example (1). Frequently, they included superlatives, also like (1). Our findings led us to formulate the tentative hypothesis that an *Es ist dies*-sentence is used if an antecedent is not easily accessible, and therefore the pronoun *dies* does not constitute an optimal candidate to be located in the prefield. The construction possibly serves to mark this circumstance pragmatically. Its use can create an effect of distance to the preceding discourse or indicate a break in the text, potentially used to take stock. These hypotheses were the starting point of our empirical investigation, which is presented in the next section.

3 Corpus Study

3.1 Method

To investigate the use of this construction, we annotated 300 *Es ist dies*-sentences randomly taken from the DeReKo corpus of written German with regard to the following categories: a) metadata, i.e. the source the sentence occurred in and the region of the source, as well as b) properties of the antecedent of *dies*, such as the distance to the antecedent measured in finite and in matrix verbs. For example, in (4) above, the distance to the antecedent of *dies* measured in finite verbs is two and one if measured in matrix verbs. Moreover, we annotated c) semantic properties of the sentences in order to account for the impression that *Es ist dies*-sentences often express evaluations or are used to take stock. There were three categories regarding the semantic properties: *It is this the nth*-constructions (5), superlatives (6) and evaluative comments (7).

4 Braunschweiger Zeitung, 22.04.2013.

(5) Es ist dies das 23. Turnier seit 1994.[5]
it is this the 23rd tournament since 1994
'This is the 23rd tournament since 1994.'

(6) Es ist dies der früheste Reisebericht über Afrika und Indien.[6]
it is this the earliest travel report about Africa and India
'This is the earliest travel report about Africa and India.'

(7) Es ist dies eine heikle und bedauerliche Entscheidung.[7]
it is this a precarious and regrettable decision
'This is a precarious and regrettable decision.'

The *Es ist dies*-sentences were compared to *Dies ist*-sentences, which can be regarded as the unmarked counterpart to the *Es ist dies*-construction and which were taken in the same proportion from the same sources as the *Es ist dies*-sentences to achieve a maximally exact mirroring.

3.2 Results

In our sample, *Es ist dies*-sentences occurred almost exclusively in texts[8] from southern regions: 38% of the instances were found in texts from Switzerland, 32% were from Austria and 18% from Bavaria. The remaining 12% were singular occurrences in texts from various regions. The results of the measurement of the distance to the antecedent are presented in Table 1. In the majority of cases, both constructions feature an antecedent located at a distance of zero finite or matrix verbs. However, *Es ist dies*-sentences refer to antecedents located at larger distances more frequently than *Dies ist*-sentences. The χ^2-test yielded significant differences between the two samples with regard to the feature 'distance to the antecedent' ($p < .01$ both for the measurement in finite verbs and for the measurement in matrix verbs).

Concerning the semantic properties of the sentences, we found 86 instances of *It is this the nth*-constructions in the *Es ist dies*-sentences opposed to 26 instances among the *Dies ist*-counterparts. The χ^2-test yielded a significant result for this difference ($p < .001$), too. The other semantic properties that were annotated, superlatives and evaluative comments, did not yield significant differences.

5 Nordkurier, 09.09.2008.
6 Tiroler Tageszeitung, 12.11.1997.
7 Süddeutsche Zeitung, 05.04.2008.
8 96% of these texts were newspaper articles.

Table 1: Distance to the antecedent (absolute values)

Distance to the antecedent	**Sentence Type**	
Number of finite verbs	*Es ist dies*	*Dies ist*
0	198	239
1	43	20
2 or more	36	17
Number of matrix verbs		
0	219	246
1	39	20
2 or more	19	10

3.3 Discussion

Speyer's ranking (2008, 2009; Table 2) predicts that prefield-*es* is only used if a sentence contains none of the preferred prefield fillers which are scene-setting elements, poset elements and topics.[9] As *dies* refers to the discourse referent the sentence is about, one could consider it to be the topic. In Speyer's prefield ranking, the notion 'topic' is defined as discourse-old and occurring in the directly preceding sentence. In the majority of cases, the *dies* of our *Es ist dies*-sentences does actually refer to an antecedent in the immediately preceding sentence. Hence, Speyer's ranking incorrectly predicts *dies*, instead of *es,* to occur in the prefield for those cases. His ranking only predicts *Es ist dies* to be the optimal candidate in cases in which *dies* refers to an antecedent that is not located in the preceding sentence.

However, *Es ist dies*-sentences are used rather rarely.[10] Hence, the fact that the *Dies ist*-version is generally more frequent and that *Es ist dies* is a marked construction should be represented in the ranking. We therefore suggest replacing

9 The first constraint 1-VF specifies that only one constituent can occur in the prefield. Scene-setting-VF requires elements such as adverbials of time to be moved to the prefield. 'Poset' stands for 'partially ordered set' and a poset relation is a type of contrast. In (5) *Fresh vegetables* and *pasta* stand in a poset relation as they are both members of the set 'food Peter buys'.

(5) Frisches Gemüse kauft Peter auf dem Markt.
fresh vegetables buys Peter at the market
'Fresh vegetables Peter buys at the supermarket.
Nudeln besorgt er immer im Supermarkt.
pasta gets he always at the supermarket
Pasta he always gets at the supermarket.'

10 To illustrate this: Our search request for *Es ist dies*-sentences yielded 4,870 hits from the DeReKo as opposed to 91,726 hits for the corresponding request for *Dies ist*-sentences.

Table 2: Speyer's (2008, 2009) prefield ranking

Candidates	1-VF	Scene-Setting-VF	Poset-VF	Topic-VF
☞ Dies ist$_{\text{Preced.Sent}}$				
Es ist dies$_{\text{Preced.Sent}}$				*
Dies ist$_{\text{NotPreced.Sent.}}$				*
☞ Es ist dies$_{\text{NotPreced.Sent}}$				

Table 3: Modified prefield ranking.

Candidates	1-VF	Scene-Setting-VF	Poset-VF	Aboutness Topic-VF	Mark-Shift-VF
☞ Dies ist$_{\text{Preced.Sent}}$					
Es ist dies$_{\text{Preced.Sent}}$				*	
☞ Dies ist$_{\text{NotPreced.Sent}}$					*
Es ist dies$_{\text{NotPrecedSent}}$				*	
☞ Dies ist$_{\text{Count}}$					*
Es ist dies$_{\text{Count}}$				*	

Topic-VF with two different constraints, AboutnessTopic-VF and MarkShift-VF (Table 3). This a first tentative approach to explain the observed phenomenon.

AboutnessTopic-VF specifies that the topic, understood as an 'aboutness' topic following Reinhart (1981), should be located in the prefield. Unlike Topic-VF, AboutnessTopic-VF does not require the topic to occur in the immediately preceding sentence which accounts for the higher frequency of *Dies ist*-sentences in general. The modified ranking specifies that, unless there is a reverse ranking, *Dies ist* is always the preferred candidate.[11] MarkShift-VF reflects the

11 In stochastic OT (Boersma & Hayes 2001) constraints are not discrete but ordered on a continuous scale of strictness. A constraint is assumed to be associated not only with one value, but with a range of values which is thought of as a probability distribution in the form of a Gaussian curve. Thus, some values have a higher probability of being selected than others. Depending on how close to each other the constraints are located on the scale, the extent to which they overlap varies. A ranking in which two constraints overlap to a large degree accounts for cases where two forms are grammatical, but one is preferred over the other. In these cases, the probability that values are selected which result in a reverse ranking is relatively high.

discourse connecting function of the prefield (see e.g. Fillipova & Strube 2007). In the default case, the prefield is expected to be filled by an element that adds to the coherence of the text. The new constraint requires a marking of breaks or unexpected moves in discourse. It has often been observed that shifts of topics tend to be marked (see Givón 1983, Bestgen/Vonk 2000, and Breindl 2008, 2011).[12] Similarly, we argue that *Es ist dies* can mark a cesura in discourse, e.g. when *dies* refers to an antecedent that is not easily accessible. We assume that MarkShift-VF slightly overlaps with AboutnessTopic-VF, which has the effect that, at times, AboutnessTopic-VF is outranked by MarkShift-VF. This accounts for the difference between *Es ist dies*-sentences and *Dies ist*-sentences with regard to the distance to the antecedent since referring back to an antecedent that is located at a greater distance is an unexpected discourse move. The significant difference in the content category *It is this the nth* is also in line with our approach as it makes sense to indicate a break in discourse when taking stock (i.e., no violation of Markshift-VF).

However, for a large number of cases, we are not yet able to pinpoint the reason for using prefield-*es*. It might be related to the often mentioned observation that anaphorically used demonstrative pronouns tend to refer to an antecedent that is harder to access (e.g. Diessel 1999: 96, Gundel et al. 2003). Using prefield-*es* could be an optional way of further highlighting this. After all, we did not find constructions such as *Es ist er* + NP oder *Es ist sie* + NP where grammatical gender already limits the number of possible antecedents.

Furthermore, an interesting question is how our modified OT ranking relates to the regional differences. We suggest that in more northern varieties of standard German the two constraints AboutnessTopic-VF and MarkShift-VF are located far apart from each other on the scale of constraints and therefore overlap to a very small extent. This has the effect that AboutnessTopic-VF outranks MarkShift-VF more regularly, which is why speakers of northern varieties of German find *Es ist dies*-sentence odd but not ungrammatical.[13] In contrast, in more southern varieties the stochastic overlap of the two constraints is stronger, which leads to a larger probability that MarkShift-VF outranks AboutnessTopic-VF. We can therefore understand regional variation as a purely stochastic difference without any modification of our proposed ranking.

12 For example, in German, one way to mark a shift of topic is inserting an adverbial in the so-called "Nacherstposition" in front of the finite verb (Breindl 2008, 2011).

13 This claim is based on the judgments from the authors of the paper as well as from other consultants from northern Germany.

4 Conclusion

The modified version of the prefield ranking incorporates a use of prefield-*es* that the old model did not factor in. It reflects that *Es ist dies*-sentences are a rarely occurring phenomenon, but it accounts for the fact that they do occur. The significant differences that were found between *Es ist dies*-sentences and their unmarked counterparts with regard to the distance to the antecedent of *dies* and the frequency of the content type *It is this the nth* were explained by the addition of the constraint MarkShift-VF. What is still needed is an explanation of those occurrences of *Es ist dies*-sentences, for which neither a large distance to the antecedent nor the content type *It is this the nt^{h}* was attested. We leave this question for further research.

References

Bestgen, Yves and Wietske Vonk. 2000. Temporal adverbials as segmentation markers in discourse comprehension. *Journal of Memory and Language* 42: 74–87.

Boersma, Paul and Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32(1): 45–86.

Breindl, Eva. 2008. Die Brigitte nun kann der Hans nicht ausstehen. Gebundene Topiks im Deutschen. *Deutsche Sprache* 36: 27–49.

Breindl, Eva. 2011. Nach Rom freilich führen viele Wege. Zur Interaktion von Informationsstruktur, Diskursstruktur und Prosodie bei der Besetzung der Nacherstposition. In Gisella Ferraresi (ed.), *Konnektoren im Deutschen und im Sprachvergleich. Beschreibung und grammatische Analyse*, 17–56. Tübingen: Narr.

Cardinaletti, Anna. 1990. Es, pro and sentential arguments in German. *Linguistische Berichte* 126: 135–164.

Diessel, Holger. 1999. *Demonstratives: Form, function and grammaticalization.* Amsterdam and Philadelphia, PA: Benjamins.

Filippova, Katja and Michael Strube. 2007. The German Vorfeld and local coherence. *Journal of Logic, Language, and Information* 16(4): 465–485.

Givón, Talmy. 1983. Topic continuity in discourse: An introduction. In Talmy Givón (ed.), *Topic continuity in discourse: A quantitative cross-language study*, 1–42. Amsterdam et al.: Benjamins.

Gundel, Jeanette K., Michael Hegarty and Kaja Borthen. 2003. Cognitive status, information structure, and pronominal reference to clausally introduced entities. *Journal of Logic and Information* 12: 281–299.

Grosz, Barbara J., Aravind K. Joshi, & Scott Weinstein. 1995. Centering: A framework for modelling local coherence of discourse. *Computational Linguistics* 21: 203–225.

Paranhos Zitterbart, Jussara. 2002. Zur Mittelfeldfähigkeit des Korrelat *es* in Verbindung mit Subjektsätzen. *Sprachwissenschaft* 27: 149–195.

Pittner, Karin and Judith Bermann. 2004. *Deutsche Syntax: Ein Arbeitsbuch*. Tübingen: Narr.

Pütz, Herbert. 1986. Über die Syntax der Pronominalform ‚es' im modernen Deutsch. Tübingen: Narr.

Reinhart, Tanya. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27(1): 53–94.

Speyer, Augustin. 2008. German vorfeld-filling as constraint interaction. In Anton Benz & Peter Kühnlein (eds.), *Constraints in discourse*, 267–290. Amsterdam and Philadelphia, PA: Benjamins.

Speyer, Augustin. 2009. Das Vorfeldranking und das Vorfeld-*es*. *Linguistische Berichte* 219: 323–353.

Tomaselli, Alessandra. 1986. Das unpersönliche „es": Eine Analyse im Rahmen der Generativen Grammatik. *Linguistische Berichte* 102: 171–190.

Walker, Marylin A., Aravind K. Joshi and Ellen F. Prince. 1998. Centering in naturally occurring discourse: An overview. In Marylin A. Walker, Aravind K. Joshi & Ellen F. Prince (eds.), *Centering Theory in Discourse*, 1–28. Oxford: Oxford University Press.

Zifonun, Gisela. 1995. Minimalia Grammaticalia: Das nicht-phorische es als Prüfstein grammatischer Theoriebildung. *Deutsche Sprache* 23: 39–60.

Zifonun, Gisela, Ludger Hoffmann, Bruno Strecker et al. 1997. *Grammatik der deutschen Sprache*, Bd. 2. Berlin: de Gruyter.

Swantje Tönnis, Lea M. Fricke, Alexander Schreiber

Methodological Considerations on Testing Argument Asymmetry in German Cleft Sentences

Abstract We present a corpus study on German *es*-clefts that tests whether subject clefts are more frequent than object clefts. This observation has been made for several other languages. However, we use a more complex method than earlier studies by not only providing the frequencies of subject/object clefts but by additionally comparing those frequencies to the general frequency of subjects/objects. Our results support the claim that subject clefts are more frequent in German. We argue that a cleft construction in its function to mark focus appears more often with subjects since there are additional options to mark focus on objects. Other features such as exhaustivity and contrast do not play a role in our cleft sample. From these results, we conclude that subjecthood is the main factor that facilitates the use of a cleft, possibly as a result of the author's intention to disambiguate focus.

Keywords German *es*-cleft, prosodic prominence, focus marking, argument asymmetry

1 Introduction

This paper presents a corpus study with the aim of contributing to a better understanding of the factors that facilitate the use of *es*-clefts in German. We analyzed crucial properties of clefts and their contexts. In this paper, we focus mainly on one aspect, namely the grammatical role of the pivot. Depending on the grammatical role of the pivot in the relative clause, we distinguish between subject clefts as in (1), and object clefts as in (2).

(1) Es war Alt-Bundespräsident Roman Herzog, der zum
It was former president Roman Herzog, who$_{NOM.SG}$ on the
50-jährigen Jubiläum eine internationale Neuorientierung
50th anniversary a international re-orientation
der Stiftung anregte.
of the foundation suggested.
'It was the former president Roman Herzog who suggested an international re-orientation of the foundation on the 50th anniversary.'
(Z07/JUL.00590 Die Zeit [Online-Ausgabe], 19.07.2007; Noble Töne, enttäuschter Nachwuchs)

(2) Es ist der Aufsteiger, den Balzac mit immer neuen
It is the climber, who$_{ACC.SG}$ Balzac with constantly new
charakterlichen Merkmalen porträtiert, […].
character features portrays, […].
'It is the (social) climber who Balzac portrays with constantly new character features.'
(R99/MAI.38158 Frankfurter Rundschau, 15.05.1999, S. 3, Ressort: ZEIT UND BILD; Zum 200. Geburtstag von Honoré de Balzac)

For several languages, it has been claimed that subject clefts are more frequent than object clefts (Carter-Thomas 2009, Roland et al. 2007, and Skopeteas & Fanselow 2010). We tested this claim for German clefts given that to our knowledge this has not been explicitly tested. Additionally, we use a more fine-grained method than earlier studies on other languages. We do not only provide the frequencies of subject and object clefts but also compare those frequencies to the general frequency of subjects and objects. It is important to take this additional step since it could be possible that subjects are just clefted more often because they are generally more frequent.

2 Background

The observation that subject clefts are more frequent than object clefts is closely related to focus marking. The cleft construction is one option for a language to realize focus, in addition to prosodic prominence, movement, and morphology. In some languages, not all of these options are equally available for all grammatical functions (see Lambrecht 2001 for French, or Hartmann & Zimmermann 2007 for West Chadic Languages). In French, for example, focus on objects can be realized via prosodic prominence, while this is not an option for subjects. According to Féry (2001), prosodic prominence is obligatorily realized at the right edge of the phonological phrase in French. Objects occur in this position and receive prosodic

prominence. Subjects, in contrast, cannot appear there. In the pivot of a cleft, however, subjects are located at the edge of a phonological phrase and receive default high prominence (see also Reinhart 1995: 62). The default intonation of a focus-background cleft[1] in French is exemplified in (3), taken from Destruel (2012).

(3) C'est BATMAN qui a pour mission d'attraper les cambrioleurs.
it-is BATMAN who has for mission to-catch the thieves.
'It is Batman who has the mission of catching thieves.'

Accordingly, Szendröi (1999: 553) proposes to analyze clefts as focus-driven movement. Similarly, DeVeaugh-Geiss et al. (2015: 386) call clefts a structural device to mark focus unambiguously. Focus on an object NP can also be realized by a cleft construction. However, there are other options for focus-marking on objects (that are inapplicable to subjects), such as default intonation and scrambling. Hence, object NPs are predicted to be clefted less often than subjects.

The aim of our study is to analyze German data with respect to the frequency of subject and object clefts and thereby gain a deeper understanding of the function of a cleft sentence. More precisely, we discuss whether the primary function of a cleft is to mark focus. German, just as French, assigns the default accent at the edge of a phonological phrase. However, it allows for more variation when it comes to intonation (see Section 4 for a detailed discussion).

3 Corpus Study

3.1 Method

We drew a random sample of 300 clefts from a sub-corpus of the DeReKo corpus[2] of written German. In our annotation, we focused on well-defined properties like the grammatical function of the cleft relative pronoun,[3] and the thematic role and animacy of the pivot NP. In order to account for the general frequency of

1 We will ignore topic-comment clefts in this paper, given that we found much more focus-background clefts in our corpus search.

2 Das Deutsche Referenzkorpus DeReKo (http://www.ids-mannheim.de/kl/projekte/korpora)/, Institut für Deutsche Sprache, Mannheim.
Since the annotation of some of the properties required a lot of context before and after the cleft sentence, we excluded texts that were not fully accessible. Moreover, we excluded *Wikipedia* articles because text coherence cannot be guaranteed due to possibly different authors for adjacent paragraphs of a text.

3 We only considered subject and object clefts. We did find some adjunct clefts where the relative pronoun was preceded by a preposition. Those clefts, however, were excluded from the analysis.

grammatical functions, we set up a comparison corpus of 200 randomly chosen non-clefted sentences from the same texts in which we found the clefts. Those sentences contained both main clauses and subordinate clauses, given that we found main clause clefts and subordinate clefts.

We analyzed the data in two ways: (i) We determined the relative frequencies of subjects and objects in the comparison corpus by counting all of their occurrences. Those frequencies were compared to the observed relative frequencies f_{cleft} of subject versus object clefts in the cleft sample. This method assumes that every grammatical argument is equally likely to be clefted, independent of the sentence it belongs to. So it ignores the fact that various grammatical arguments are unevenly distributed in sentences. (ii) For the second analysis, it is assumed that each sentence is equally likely to become a cleft. As sentences can have different numbers of arguments, this means that arguments of different sentences can now have different probabilities to be clefted. For example, compare two sentences of the form S-V-O and S-V-O-O. Both sentences are equally likely to become clefts, but have a different number of grammatical arguments. If the first sentence is selected, the probability that the subject is clefted is 0.5, as there are only two grammatical arguments which can be clefted. It is not possible to cleft the verb. If the second sentence is selected, this probability drops to 0.33, as there are now three possible grammatical arguments to be clefted. We calculated the probability of being clefted for each subject and object in each sentence from the comparison corpus and calculated their average over all sentences p_{cleft}, which was then compared to f_{cleft}. Each of the approaches can be seen as a useful simplification because the aspects they ignore are independent of each other.

We annotated several other properties of each cleft and its context. It is generally assumed in literature that clefts have an existence presupposition and an exhaustivity inference of some sort. The following inferences would be predicted for the cleft in (1).

a. Existence presupposition: Somebody suggested an international re-orientation of the foundation on the 50th anniversary.
b. Exhaustivity inference: Nobody other than the former president Roman Herzog suggested an international re-orientation of the foundation on the 50th anniversary.

The analysis of those inferences in our corpus, however, turned out to be unfeasible, as the inter-annotator agreement was too low for these features. The property ‘contrast’ was especially difficult to annotate since the notion is not well-defined. Following Repp (2010), we did annotate some categories related to contrast, such as the existence of explicitly mentioned alternatives and their

negation. Those categories did not seem to play a role in our sample. Taking an intuitive point of view, however, the cleft in example (4) clearly constitutes a contrast between Tony Blair and Gordon Brown. For the purpose of annotation, however, it is not obvious as to how to operationalize contrast in this and similar examples. Repp's (2010) criteria do not apply.

(4) Tony Blair, zuletzt in seiner Partei geradezu verhasst, hat sensationelle drei
Tony Blair, lately in his party virtually hated, has sensational three
Wahlsiege errungen; es war sein nach links rückender Nachfolger
election victories achieved; it was his to left moving successor
Gordon Brown, der abgewählt wurde.
Gordon Brown, who voted out was.
'Tony Blair, who was virtually hated in his party lately, achieved three sensational election victories; it was his left-moving successor Gordon Brown who was voted out of office.'
(Z10/OKT.03679 Die Zeit [Online-Ausgabe], 07.10.2010; Abschied vom Klassenfeind)

Since these properties did not seem to play a role in our sample, we will ignore them in our analysis.

3.2 Results

Table 1 presents the absolute numbers of subjects and objects found in the comparison sample and in the cleft pivots of the cleft sample.

Table 1: Absolute numbers n_{cleft} for the cleft sample and n_{comp} for the comparison corpus.

	n_{cleft}	n_{comp}
Subjects	249	192
Objects	24	93

Both approaches described above yield that subject clefts occur significantly more often than object clefts even with respect to the general frequency of subjects and objects. For approach (i), we tested the relative frequencies f_{cleft} of subjects and objects from the cleft sample and the relative frequencies f_{comp} from the comparison corpus for significant deviation using a χ^2-test. The frequencies are displayed in Table 2. The test shows that subject clefts are significantly more frequent in the cleft sample ($p < 0.01$).

Table 2: Frequencies of subjects and objects in the cleft sample (f_{cleft}) and the comparison sample (f_{comp}), and the average probability (p_{cleft}) of subjects and objects in the comparison sample.

	f_{cleft}	f_{comp}	p_{cleft}
Subjects	0.91	0.67	0.76
Objects	0.09	0.33	0.24

For approach (ii), we tested f_{cleft} and the average probabilities p_{cleft} of subjects and objects from the comparison corpus (also displayed in Table 2) for significant deviation using a t-test. This test shows that subject clefts are significantly more frequent in the cleft sample than predicted by p_{cleft} ($p < 0.01$).

One natural explanation of the data could be that subjects are just clefted more frequently because of other properties that often co-occur with subjecthood, such as agentivity and animacy. After comparing these properties for subjects in the comparison corpus und subjects in the cleft pivots of our cleft sample, we can rule out this objection. Table 3 and 4 show that both samples demonstrate the same distribution for animate/non-animate and agentive/non-agentive subjects. A χ^2-test yielded a p-value of $p = 0.39$ for animacy and $p=0.56$ for agentivity. Hence, those properties do not seem to be the crucial ones.

Table 3: Absolute numbers (and %) of (in-) animate subjects in the cleft sample n_{cleft} and the comparison corpus n_{comp}.

	n_{cleft}	n_{comp}
Subjects [+animate]	117 (47%)	97 (52%)
Subjects [-animate]	132 (53%)	91 (48%)

Table 4: Absolute numbers (and %) of (non-) agentive subjects in the cleft sample n_{cleft} and the comparison corpus n_{comp}.

	n_{cleft}	n_{comp}
Subjects [+agent]	81 (33%)	71 (37%)
Subjects [-agent]	161 (67%)	123 (63%)

4 Discussion

Our results indicate a higher frequency of subject clefts as opposed to object clefts in German. We follow a line of argumentation similar to what is proposed by Féry (2001) and Szendröi (1999). We take focus to be a semantic notion (Krifka 2008). The focused element is syntactically marked by an F-feature (Rooth 1992) which is realized at the phonological form with an A-accent (Bolinger 1958). Contrary to French, in spoken German it is generally possible to mark focus by intonation in any position (including the subject position), as indicated in (5). However, this is different when it comes to written German. Here, the reader cannot identify the focus by referring to intonation but needs to rely on other cues provided in the text. The overt question in (5) could be such a cue.

(5) Wer hat einen Apfel gegessen? – NINA hat einen Apfel gegessen.
Who has an apple eaten? – NINA has an apple eaten.
'Who ate an apple? – NINA ate an apple.'

(6) Nina hat einen APFEL gegessen.
Nina has an APPLE eaten.
'Nina ate an APPLE.'

(7) Es ist NINA, die einen Apfel gegessen hat.
It is NINA who an apple eaten has.
'It is NINA who ate an apple.'

(8) Nina hat das Buch dem MANN geschenkt.
Nina has the book the MAN given.
'Nina gave the book to the MAN.'

If the context does not provide such a cue, the reader is likely to rely on her knowledge of where the default focus accent lies, that is, as in French, at the right edge of a phonological phrase.[4] In many cases, the default intonation results in the object (not the subject) receiving highest prominence, as in (6). Hence, the object would be identified as the focus. Furthermore, objects can be scrambled into a position where they receive the default focus accent. In (8),

4 This does not imply that the reader constructs an actual prosodic-phonological representation for the written text although some studies would support that (for an overview of related research see Leinenger 2014). For our argument to hold, it suffices that the reader just uses her knowledge of where the accent is 'usually' assigned.

for instance, the indirect object NP *dem Mann* ('the man') is scrambled to the end of the phonological phrase, where it is focused by default. In order to disambiguate focus-marking on the subject in written German, special marking is helpful (DeVeaugh-Geiss et al. 2015). The cleft construction puts the subject into a position where it receives highest prominence by default (Szendröi 1999) and, thus, gives the reader a cue to identify the subject as the focus (see example (7)).

Following Féry (2001) and DeVeaugh-Geiss et al. (2015), we argue that a cleft construction in its function of marking focus appears more often with subjects since there are other additional options to mark focus on objects, such as default focus accent or scrambling, which are inapplicable to subjects. In their base position, subjects do normally not receive a default accent. Furthermore, subjects are unlikely to be scrambled in order to be focused.

The question is now whether disambiguating subject focus is indeed the main motivation for using a cleft. Literature on clefts has mentioned several other features of clefts that might be worth considering, e.g., exhaustivity or the existential presupposition as explained in Section 3.1. Firstly, our annotation data did not provide clear evidence for the relevance of those features. Our argument is further strengthened by the observation that clefts are hardly ever used in spoken German.[5] An account just based on the existential presupposition and/or exhaustivity cannot explain the difference between the frequency of clefts in spoken and written German. Neither the existential presupposition nor the exhaustivity inference seem plausible to have an effect on the frequency of clefts in general. In particular, there is no reason why those properties should be more developed in written than in spoken German.

Our analysis of clefts as devices to shift prominence away from the default, in contrast, predicts there to be fewer clefts in spoken German. In spoken German there is simply no need for a cleft construction since focus can always be disambiguated using intonation by marking an element in-situ, as in (5). This option is missing in written German, which leads to more clefts in written German. Our analysis is nevertheless compatible with assigning an exhaustivity inference and an existential presupposition to clefts, but those features are not assumed to constitute the main motivation for using a cleft.

5 Even though we did not conduct a quantitative study about the frequency of clefts in spoken German, our informants and the native speaker judgments of the authors support the low frequency of clefts in spoken German.

5 Conclusion

From our data set, we can conclude that subjecthood is the main factor determining the use of clefts, possibly due to the wish of the author to give cues for unambiguously identifying the focused element in the sentence. This is in line with the observation that subject clefts occur more often than object clefts since German has other ways of disambiguating focus for objects, e.g., default intonation and scrambling. Our approach is also capable of predicting a difference between spoken and written German.

Some issues are left open here and will need further research. So far, our reasoning only works for focus-background clefts, but should be extended to also cover topic-comment clefts. Moreover, the role of contrast should be operationalized for annotation or further analyzed using other methods.

References

Bolinger, Dwight L. 1958. A Theory of Pitch Accent in English. *Word*, 14(2–3), 109–149.

Carter-Thomas, Shirley 2009. The French c'est-cleft: Function and frequency. In D. Banks (ed.), *La linguistique systémique fonctionnelle et la langue française*, 127–157. Paris : L'Harmattan.

Destruel, Emilie. 2012. The French c'est-cleft: An empirical study on its meaning and use. *Empirical issues in Syntax and Semantics*, 9, 95–112.

DeVeaugh-Geiss, Joseph P., Malte Zimmermann, Edgar Onea and Anna-Christina Boell. 2015. Contradicting (not-)at-issueness in exclusives and clefts: An empirical study. *Semantics and Linguistic Theory*, 25, 373–393.

Féry, Caroline. 2001. Focus and phrasing in French. In C. Féry and W. Sternefeld (eds.), *Audiatur Vox Sapientia. A Festschrift for Arnim von Stechow*, 153–181. Berlin: Akademie Verlag.

Hartmann, Katharina & Zimmermann, Malte. 2007. In place-out of place? Focus in Hausa. In K. Schwabe and S. Winkler (eds.), *On Information Structure, Meaning and Form: Generalizations across languages*, 365–403. Amsterdam: John Benjamins Publishing Company.

Krifka, Manfred. 2008. Basic notions of information structure. *Acta Linguistica Hungarica*, 55(3–4): 243–276.

Lambrecht, Knut 2001. A framework for the analysis of cleft constructions. *Linguistics*, 39(3): 463–516.

Leinenger, Mallorie. 2014. Phonological coding during reading. *Psychological Bulletin*, 140(6): 1534–1555.

Reinhart, Tanya 1995. Interface strategies. *OTS working papers in linguistics*.

Repp, Sophie 2010. Defining 'contrast' as an information-structural notion in grammar. *Lingua*, 120(6): 1333–1345.

Roland, Douglas, Frederic Dick and Jeffrey L. Elman. 2007. Frequency of basic English grammatical structures: A corpus analysis. *Journal of Memory and Language*, 57(3): 348–379.

Rooth, Mats. 1992. A theory of focus interpretation. *Natural language semantics*, 1(1): 75–116.

Skopeteas, Stavros and Gisbert Fanselow. 2010. Focus types and argument asymmetries: A crosslinguistic study in language production. In C. Breul and E. Göbbel (eds.), *Contrastive information structure*, 169–197. Amsterdam: John Benjamins Publishing Company.

Szendröi, Kriszta. 1999. A stress-driven approach to the syntax of focus. *UCL Working Papers in Linguistics*, 11, 545–573.

Johanna Marie Poppek, Tibor Kiss, Francis Jeffry Pelletier

Kinds, Containers, Instances: Mass Nouns and Plurality

Abstract The existence of formally realized plurality in the domain of mass nouns is a major challenge, especially if the hypothesis is taken that mass nouns possess some kind of "built in" plurality as their main distinguishing feature compared to count nouns. To address this issue, we performed a large-scale corpus study on the plural occurrences of mass nouns and dual life nouns using the OANC corpus and a database of noun-sense pairs annotated in terms of their countability class. Results showed that not only do pluralizations of mass terms occur frequently in the corpus, the nature of their meaning shifts differs with regards to their specific countability class, providing a deeper insight into the semantic and pragmatic nature of the count and mass continuum.

Keywords Countability, plural, mass terms, corpus study

1 Introduction

The existence of formally realized plurality in the domain of mass nouns is a major challenge, especially if the hypothesis is taken that mass nouns possess some kind of "built in" plurality as their main distinguishing feature compared to count nouns, noting that nouns that possess a (morphological) plural are usually considered count (e.g., Chierchia 1998). Other approaches stress the general similarity of mass nouns and plural expressions, leaving out the field of plurality *of* mass nouns (e.g., Lasersohn 2011).

In this article, we will present a large-scale corpus study as an approach for a systematic analysis of mass terms and plurality and their implications. It is based on a fine-grained nominal classification resource (*Bochum English Countability Lexicon* Kiss et. al. 2014 and 2016) that eschews both a binary distinction and a lemma-based approach to countability.

1.1 Data

Since the analysis of the general phenomenon of countability is usually described as a binary feature dividing the domain into only two realms of countable and uncountable nouns (e.g., Borer 2005) and only addressed with a small set of staple nouns, we created a database that allows the study of countability on a larger scale and in a more fine-grained way. The database consists of approximately 12,000 English noun-sense pairs that were enriched with their WordNet definitions for every sense and annotated in terms of countability by four native speaker annotators using a set of six pattern test questions to test their semantic and syntactic behavior. The resulting 18 subclasses are grouped in four major classes that represent the general complexity of the countability issue. Table 1 shows the general distribution of major classes for the resulting pairs consistently annotated by at least two annotators and provides examples for each class in terms of WordNet lemma, POS-Tag and sense number. Note that the names of the subclasses are an artifact of the initial classification process carried out in R and were kept as neutral captions for the respective classes.

It should be stressed here that the annotation for every sense did apply at the type level without any access to corpus data, not at the token level. Therefore, a classification e.g., as *both mass and count* does imply a deviant position in the count-mass continuum or a dual life nature, while *neither mass nor count* contains senses where the whole distinction does not seem to apply, e.g., unique entities. It should further be noted that although most noun-sense pairs are classified as *regular count* or *regular mass* as accounted for in the literature, there is a relevant amount of data that does not fit into the binary scheme, showing that the issue of countability resembles more a continuum or a spectrum than a distinction.

Table 1: Major Classes of BECL

Major Class	Frequency	Subclasses	Examples
Regular Count	8,434	235, 721, 371, 73	*animal.n.01*; *childhood.n.01*; *manners.n.01*; *making.n.03*
Regular Mass	2,427	528, 519, 531	*knowledge.n.01*; *adaptability.n.01*; *lingerie.n.01*
Both Mass and Count	699	510, 726, 729, 513	*glue.n.01; superstition.n.01; theft.n.01*; *china.n.04*
Neither Mass nor Count	315	523, 37, 190, 514, 199, 28, 353	*doomsday.n.02; infinite.n.01; midline.n.01; provenance.n.01; heyday.n.01; midst.n.01; hamlet.n.02*
Total	11,875	18	

To further support research on the issue, the resource is made publicly available via http://www.count-and-mass.org.

1.2 Approach

Although the nominal classification of the resource in terms of countability already allowed insight into the type level of countability, there are several approaches to address the phenomenon of mass-to-count or count-to-mass shifts at the *token level* (e.g.; De Belder 2008b; Nicholas 2002 among others) Mass-to-count shifts are usually determined by a "deviant" behavior of a noun that is usually classified as mass. "Deviant behavior", in case of mass terms, could occur with an indefinite article in the singular or with a morphologically realized plural. This does not include cases like pluralia tantum (e. g., *scissors*) that fall into a different countability category, but rather to genuine mass terms that occur in a plural form.

To determine the distribution of plural occurrences of apparent mass nouns, we have used the Stanford NLP system[1] to parse sentences from the *Open American National Corpus* (OANC, *http://www.anc.org*) containing nouns from three mass noun classes of the database (528, 510 and 726) and extracted sentences that showed plural occurrences despite the nouns being classified as *mass nouns exclusively* (528) and *dual use nouns* (510, 726). For more information on the classes and their annotation pattern cf. Table 2.

2 Corpus Study on Plural Mass Terms

The phenomenon addressed here takes place at the token level of a specific lemma. Since our data is annotated at the sense level, we took only completely annotated lemmata into account (meaning that all senses of a lemma that WordNet provides must be present in our data) that consistently belong to one subclass with respect to all their senses. The general hypothesis is that mass nouns of class 528 should not possess a morphological plural, while plural occurrences of mass terms from class 510 and 726 should be accompanied by a meaning shift (cf. Borer 2005; Chierchia 1998 on plural meaning shifts on mass terms).

The sentences extracted from the OANC corpus contained approximately 1,900 plurality examples for class 528 (167 lemmata), approximately 5,400 examples for class 510 (241 lemmata) and approximately 1,500 plural occurrences (64 lemmata) for class 726. Most lemmata contained in all three classes showed

1 Included in the parser software package *(http://nlp.stanford.edu/software/nndep.shtml).*

Table 2: Annotation Patterns

528 (regular mass)	510 (both mass and count)	726 (both mass and count)
Can be combined with *more*, the resulting sentence uses a mode of measurement other than number	Can be combined with *more*, the resulting sentence uses a mode of measurement other than number	Can be combined with *more*, the resulting sentence uses a mode of measurement other than number
more + morphological plural is *not applicable*	*more* + morphological plural is possible and semantically equivalent to a sentence with an explicit classifier	*more* + morphological plural is possible and semantically equivalent to a sentence with an explicit classifier
Singular form can be subject of a classification or definition without, but not with an indefinite determiner (*A <sense> is a kind of X)	Singular form can be subject of a classification or definition without, but not with an indefinite determiner (*A <sense> is a kind of X)	Singular form can be subject of a classification or definition with and without an indefinite determiner (A <sense> is a kind of X)
Example: *flexibility*	Example: *punishment*	Example: *friendship*

several plural occurrences in the corpus, so we can assume that mass plurals are not a rare phenomenon. Besides these generally high frequencies, all three classes showed a behavior that can be described as mass-to-count type shifting. Type shifting, for this matter, would indicate an arising interpretation as a *kind*, a *unit* or an *instantiation* of an act, event or result (cf. Table 3 for examples from OANC).

Table 3: Type Shifting Examples

Unit Interpretation: Three carboxy-terminal tyrosines (positions 624-6), hypothesized to play regulatory roles, were replaced by *phenylalanines*.
Kind Interpretation: The universe, in short, is breaking *symmetries* all the time by generating such novelties, creating distinctive molecules or other forms which had never existed before.
Instantiation Interpretation: The reaction products were purified by means of three repeated gel *chromatographies* using water-saturated Sephadex G-50 in Millipore/ Multiscreen filtration plates according to the instructions provided by the supplier and dried under vacuum.

Kind interpretations could also be labelled *species interpretation* or *type interpretation* and imply an interpretation shift from a (bare) mass reading towards an element from a greater variety, meaning a *kind of something* or creating a *class of objects*. By this process of interpretation, a mass term can obtain a countable interpretation.

(1) Experts tend to implicate increased environmental exposure to *carcinogens*.

Unit interpretations describe the general phenomenon of *containering* a mass term into certain bits or quantities thus allowing it to be counted, usually by using a specific measure phrase. However, the corpus study showed that unit interpretations do not require a measure phrase, but can also be contextually derived:

(2) The PCR products of the ITS were resolved as single bands on 1 % agarose *gels*. (without measure phrase)

(3) Using an experimental group and a control group, researchers would compare levels of *pesticides* found in settled dust, on children's hands, and in their blood, urine, or hair. (with measure phrase)

It should be noted here that especially unit interpretations without a specific measure phrase can be easily confused with kind interpretations. Since the example (2) provides contextual information that the plural of *gel* refers to the same type of object, the plural is interpreted as *portions* of something, not *kinds*, in contrast to example (1) where the context strongly suggests a number of different kinds of *carcinogens*. Nonetheless, both categories show a certain amount of overlap in some cases depending on the nature of the noun.

We observed another kind of type shift we call the *instantiation interpretation*. In those cases, nouns are coerced into a countable noun by an interpretation as an act, an event or a result.

(4) In most places, heavy *snowfalls* are considered a troublesome (albeit picturesque) natural phenomenon.

This type of mass-to-count shifting is rarely discussed in the literature, and when it is, it is usually described as a restricted extension to certain categories of nouns, and as neither regular nor predictable (e.g., by Payne & Huddleston 2002).

While those kinds of mass term pluralization have been partly described in the literature (cf. e.g., De Belder 2008a and 2008b or Payne & Huddleston 2002), the phenomenon has, to our knowledge, not been addressed on a large scale in terms of observing general frequencies and implications of plural mass terms.

Although all three classes showed all three kinds of meaning shifts, the distribution of shifting interpretations strongly differs, resulting in a stronger preference for a *unit interpretation* or an *instantiation interpretation* for dual use nouns and as an *instantiation* or a *kind* for proper mass nouns.

These empirical results provide impulses for two observations. First, mass term plurals seem to occur with regularity and show a certain variation which is

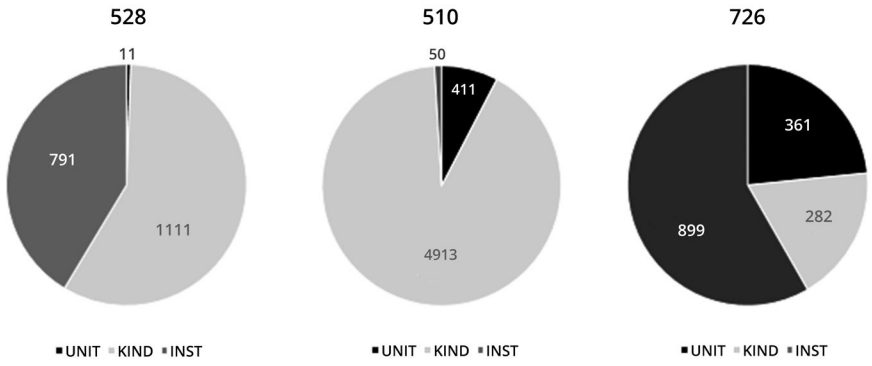


Figure 1: Type-Shifting Distribution.

greater than that which is accounted for by other researchers. In particular, the frequent observation of instantiation type-shifting formerly classified as rather rare (cf. Payne & Huddleston 2002) implies that the phenomenon might be a lot more common than hitherto thought. Second, the countability subclass seems to have a strong effect on the general distribution of the type-shifting classes (and also their frequency), implying that there might be a semantic effect that is revealed though a large-scale analysis (Figure 1).

3 Conclusion and Further Work

This first corpus study showed that although they are neglected by a large amount of current research, mass term plurals frequently occur in actual language data. In addition to this, our observations imply that they also follow certain regularities. Since the kind of type shift also seems to be based on the general semantic nature of the noun (only abstract nouns can undergo an instantiation type shift, for example, cf. Payne & Huddleston 2002), the general distribution of the coercion examples also allows a closer look at a general semantic pattern that might influence the position of a noun inside the countability continuum and to clarify to what extent those phenomena could be the result of a systematic polysemy. The variation inside the data of the different subclasses also implies that a more fine-grained view of the count and mass spectrum can provide a deeper insight into mechanisms that might be overlooked in a broader classification.

The data extracted thus provides the basis for an account of the varying effects of plurality within the class of "mass terms" and shows how large-scale corpus studies are able to address a basically underresourced phenomenon. Further research will extend to similar countability classes as well as analyzing the general semantic and pragmatic nature of pluralization of mass nouns.

References

Alexiadou, Artemis. 2011. Plural Mass Nouns and the Morphosyntax of Number. In *Proceedings of the 28th West Coast Conference on Formal Linguistics,* 33–41.

Borer, Hagit. 2005. *Structuring Sense. Vol. I: In Name Only*. Oxford: Oxford University Press.

Chierchia, Gennaro. 1998. Plurality of Mass Nouns and the Notion of the "Semantic Parameter". In Susan Rothstein (ed.), *Events in Grammar*. Dordrecht: Kluwer, 53–103.

De Belder, Marijke. 2008a. Size matters: Towards a syntactic decomposition of countability. In Natasha Abner and Jason Bishop (eds.), *Proceedings of the 27th West Coast Conference on Formal Linguistics*. Somerville: Cascadilla Proceedings Project.

De Belder, Marijke. 2008b. Sizing up countability*:* Towards a more fine-grained mass-count distinction. Talk at ConSOLE, Paris, 10-12 January 2008. https://lirias.kuleuven.be/bitstream/123456789/409023/1/abstract+Console+De+Belder+(%2Bname).pdf.

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Katz, Graham and Roberto Zamparelli. 2011. Meaning-shifting plurality and the Count/Mass Distinction. In: *QITL-4 – Proceedings of Quantitative Investigations in Theoretical Linguistics* 4 29.03.2011–31.03.2011, Berlin, Humboldt-Universität, 43–46.

Kiss, Tibor, Francis Jeffry Pelletier, Halima Husic, Johanna Marie Poppek and R. Nino Simunic. 2016. A Sense-Based Lexicon for Count and Mass Expressions: The Bochum English Countability Lexicon. In *Proceedings of LREC 2016*. Portorož, Slowenia.

Kiss, Tibor, Francis Jeffry Pelletier and Tobias Stadtfeld. 2014. Building a Reference Lexicon for Countability in English. In *Proceedings of LREC 2014*. Reykjavik, Iceland.

Lasersohn, Peter. 2011. Mass Nouns and Plurals. In: *Semantics: An International Handbook of Natural Language Meaning*. Berlin: de Gruyter.

Nicolas, David. 2002. Conversions of count nouns into mass nouns in French: the roles of semantic and pragmatic factors in their interpretations. https://halshs.archives-ouvertes.fr/ijn_00000623v2/document.

Nicolas, David. 2008. Mass nouns and plural logic. In *Linguistics and Philosophy*, 31 (2): 211–244.

Payne, John, Rodney Huddleston 2002. Nouns and Noun Phrases. In *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press. 323–524.

Stefan Heck

Verbal Aspect in the Czech and Russian Imperative

Abstract The opposing perfective (PV) and imperfective (IPV) aspects are not used uniformly across Slavic languages. One of the areas of variation is the imperative, where especially Russian is known to express special pragmatic meanings (politeness and rudeness) through the IPV (Padučeva ²2010, Benacchio 2010), a possibility which other languages like Czech possibly lack. Using corpus data, this paper attempts to check Benacchio's claims that Czech makes almost no use of the pragmatic IPV imperative. One study compares the relative frequencies of PV and IPV imperatives for a chosen number of aspect pairs in Czech, Polish and Russian using the *Aranea* webcorpora; the other study uses the parallel corpus *InterCorp* (v9) to compare the frequency of Czech IPV imperatives corresponding to Russian PV and vice versa. Both studies show the IPV imperative to be more widespread in Russian than in Czech (and Polish), lending support to Benacchio's claims.

Keywords Verbal aspect, imperative, politeness, parallel corpus, Slavic

1 Introduction[1]

From a morphological point of view, Slavic aspect is expressed derivationally. Aspectual verb pairs like Polish imperfective (IPV) *robić* and perfective (PV) *zrobić* 'do', or Czech PV *odhalit* and IPV *odhalovat* 'reveal', are formed using a number of derivational affixes, sometimes with slight changes to the verb stem, and in some few cases with suppletive forms. Thanks to this, the aspectual opposition permeates almost the entire verbal paradigm including participles, the infinitive, and the imperative. While the inventory of aspectual morphology is

1 I wish to thank my anonymous reviewers for their kind and helpful comments to my initial manuscript. Any shortcomings of the present paper remain, of course, entirely my own fault.

remarkably similar across Slavic languages, the way in which the grammemes [PV] and [IPV] are employed in certain domains (iteration, performative speech acts, Historical Present etc.) shows considerable inner-Slavic variation. Eckert (1984) and Stunová (1993) for example, compare Czech and Russian aspect, and Dickey (2000) compares aspect in all major Slavic languages for several phenomena (but not the imperative), concluding that Slavic aspect use can be divided roughly into an Eastern type (Russian, Ukrainian, Belarusian, Bulgarian and Macedonian) and a Western type (Czech, Slovak, Slovenian and the Sorbian languages), with Polish and Bosnian/Croatian/Serbian in a transitional zone.

1.1 Standard aspect use in the imperative

Aspect use in the imperative has been described among others by Padučeva (²2010), Lehmann (2008), Wiemer (2008), and by Benacchio (2010) in a comparative monograph comprising all major Slavic standard languages. In general, aspect use in the imperative follows what may be called "canonic" aspect functions as they are described e.g. in the AG-80 or Lehmann (2009) for Russian or in Dickey (2000) for Slavic languages in general: the PV is used for achievements and accomplishments, the IPV for states and activities. Cf. the following PV examples:

(1) a. Otevři dveře, prosím! (Cz)
b. Otwórz drzwi, proszę! (Pl)
c. Otkroj dver', požalujsta! (Ru)
open.PV.IMP.SG door please
'Please open the door!' (Benacchio 2000: 80)

The IPV is used also in open iterations, as in general advice, cf. the following:

(2) Chladničku otevírejte vždy pouze na krátkou dobu. (Cz)
freezer open.IPV.IMP.PL always only on short time
'Always open the freezer for a short time only!' (SYN2015)

(3) Kupuj zawsze u mnie. (Pl)
buy.IPV.IMP.SG always at me
'Always buy from me.' (NKJP)

(4) Pokupaj, poka deševle. (Ru)
buy.IPV.IMP.SG as-long-as cheap.COMP
'Buy while it's cheaper.' (NKRJa)

Note that Czech and Russian differ in that Czech also allows for the PV in iteration, being able to focus on the perfective micro-event rather than the macro-level of iteration, whereas Russian allows the PV only in the so-called summary meaning (Stunová 1993, Dübbers 2015).

Finally, the IPV is also regularly used under negation, the negated PV being possible only in non-volitional contexts (cf. Wiemer 2001 or Lehmann 2009).

Up to this point, aspect usage in the imperative has not been very surprising. Let us now turn to a new set of examples.

1.2 The pragmatic use of aspect: politeness/rudeness

(5) a. Segodnja na ulice xolodno, oden'tes'.PV teplee. (Ru)
b. Segodnja na ulice xolodno, odevajtes'.IPV teplee.
today on street cold dress.IMP.PL warm.COMP
'It is cold outside today, dress warmer.' (Benacchio 2010: 50)

(6) a. Pokažite.PV dokumenty! (Ru)
b. Pokazyvajte.IPV dokumenty!
show.IMP.PL documents
'Show your documents!' (Benacchio 2010: 51)

In (5) and (6), Russian allows the use of the IPV aspect although the situation described by the verb is neither an activity, nor iterated, nor negated. The PV is just as possible in this context. The IPV is said to make the statement more soft and polite in (5), more rude in (6). These pragmatic effects of politeness/rudeness are the focus of Benacchio (2010) and are well known and described for Russian (cf. also Padučeva ²2010, Wiemer 2008, Lehmann 2008). Both Padučeva and Benacchio also explain, in different ways, how the effects of positive politeness vs. rudeness arise in the situational context. Whether this pragmatic use of the IPV is also found in Polish or Czech is less clear. Eckert (1984) notes that the IPV is used in "certain standard etiquette forms of polite address" and also "to add politeness to an order expressed by verbs rendering a concrete movement" (139) in Russian, but not in Czech, that is to say, she acknowledges a pragmatic difference, but does not point out the possible rudeness of the IPV. According to Benacchio (2010), the positive-politeness effect is completely unavailable in Czech and the rudeness effect is also very limited, possibly exclusive to substandard language, while in Polish both are possible, but still more limited than in Russian. The exact nature of these limitations is not clear.

This is where this paper comes in. I conducted two studies to test Benacchio's informant-based claims against corpora, more specifically, to find out whether

Russian really uses the IPV imperative more than Polish, and Polish in turn more than Czech.

There has been a previous corpus study on aspect in the Slavic imperative by von Waldenfels (2012), who analysed 11 Slavic languages in his parallel corpus ParaSol. He calculated and visualised distances between the individual languages, however his study considered only whether languages differed or not for each imperative in the text, but not in which way (i.e. a Czech IPV corresponding to a Russian PV imperative was not distinguished from a Cz.PV-Ru.IPV pairing), so this is of little help here.

2 Corpus study #1: Comparison of the frequency of IPV and PV partners in the imperatives

This study was conducted using Vladimír Benko's *Aranea* webcorpora in Czech (Araneum Bohemicum Maius 15.04), Polish (Araneum Polonicum Maius 15.02) and Russian (Araneum Russicum Maius 15.02).

I extracted the frequencies of occurrence of IPV and PV non-negated imperatives by lemma, paired the aspectual partners together and calculated the percentage of how many imperatives of the given aspect pair are IPV. These are given in table 1. For reference, I added the percentage of IPV tokens for each aspectual pair from the entire lexemes. For the sake of brevity and for ease of comparison, I will discuss only the singular here: while Czech and Russian both have a simple T-V-distinction for formality comparable to French, Polish uses a system of address nouns (*pan* 'sir', *pani* 'madam', *państwo* for mixed groups, among others) with third-person agreement. This means that a Czech 2PL imperative like *dejte!* 'give!' can have several Polish equivalents, depending on who exactly is addressed, which complicates comparison between these languages.

The image is not as clear-cut as we might have hoped, but our predictions are at least confirmed. Consider, for example, the equivalents of English *to sit down*: Czech makes almost exclusive use of the PV while Polish and Russian use the IPV as well, with Russian even favoring it. The situation for *to look (at)* is very similar, and to a lesser extent for others as well. In some cases, Polish appears closer to Czech than Russian, in other cases it even uses less IPV than Czech does. For some verbs, there is no discernible difference between languages (e.g. *to allow, to stop, to try*), while for *to help* and *to ask* Czech surprisingly has the highest percentage of IPV imperatives.

The high percentage of Czech IPV *dávat* 'give' surprises at first glance, but about half of these cases (745 out of 1423) are part of the phraseologism *dávat (si) pozor/dávat (si) bacha* 'to be careful'. In these cases the IPV is perfectly natural.

Table 1: Percentages of IPV partner verbs in singular imperatives vs. for all wordforms.

English	CZ PV	CZ IPV	% IPV in imp	% IPV total	PL PV	PL IPV	% IPV in imp	% IPV total	RU PV	RU IPV	% IPV in imp	% IPV total
allow	dovolit	dovolovat	0.6	23.1	pozwolić	pozwalać	2.3	61.7	razrešit’	razrešať	3.1	23.4
ask	zeptat	ptát	29.4	42.6	s-, zapytać	pytać	15.2	45.7	sprosit’	sprašivať	12.8	38.9
close	zavřít	zavírat	12.5	21.1	zamknąć	zamykać	2.4	30.1	zakryt’	zakryvať	13.7	27.5
do	udělat	dělat	45.9	55.8	zrobić	robić	24.1	49.8	sdelať	delať	50.1	47.7
find	najít	nacházet	0.9	29.6	znaleźć	znajdować	0.0	42.8	najti	naxodiť	3.5	16.0
give	dát	dávat	48.2	19.8	dać	dawać	10.0	45.1	dať	davať	20.8	46.9
help	pomoct	pomáhat	24.7	34.1	pomóc	pomagać	5.2	38.8	pomoč’	pomogať	9.7	36.3
listen	poslechnout	poslouchat	77.3	71.9	posłuchać	słuchać	55.5	80.8	poslušať	slušať	66.2	78.3
look	podívat	dívat	10.7	27.1	popatrzeć	patrzeć	84.4	90.1	posmotreť	smotreť	92.2	63.8
pay	zaplatit	platit	20.3	73.4	zapłacić	płacić	20.4	58.1	zaplatiť	platiť	74.6	70.9
remain	zůstat	zůstávat	2.9	32.1	zostać	zostawać	0.0	6.5	ostat’sja	ostavat’sja	61.1	47.2
return	vrátit	vracet	1.3	29.6	wrócić	wracać	34.4	48.4	vernut’sja, vozvratit’sja	vozvraščat’sja	37.2	30.6
say, tell	říci	říkat	8.4	38.7	powiedzieć	mówić	22.0	57.9	skazať	govoriť	24.1	53.2
id.	povědět	povídat	42.3	70.0	–	–	–	–	–	–	–	–
show	ukázat	ukazovat	2.0	35.5	pokazać	pokazywać	3.3	45.9	pokazať	pokazyvať	9.9	36.3
sit down	sednout	sedat	1.8	10.3	usiąść	siadać	34.2	41.0	sesť	sadit’sja	68.9	27.5
id.	posadit	posazovat	0.3	1.1	–	–	–	–	–	–	–	–
stop	přestat	přestávat	0.0	16.8	przestać	przestawać	0.0	28.2	perestať	perestavať	0.0	28.0
take	vzít	brát	1.1	43.8	wziąć	brać	13.9	50.5	vzjať	brať	32.0	34.6
try	zkusit	zkoušet	2.5	31.4	spróbować	próbować	6.4	63.9	poprobovať	probovať	4.7	22.9
wait	počkat	čekat	8.6	87.5	poczekać	czekać	47.4	91.3	podoždať	ždať	52.3	91.3
write	napsat	psát	26.9	50.9	napisać	pisać	24.4	59.5	napisať	pisať	58.2	57.7

This leads us to an important point: the respective verbs in one row of the above table are of course not perfect equivalents. To illustrate this on *to give*, Czech PV *dát* is often used as 'to put' as in *dej to na stůl* 'put this on the table', and the Russian IPV imperative *davaj/davajte* is often used with an exhortative meaning, to such an extent that it can be described as a particle meaning 'come on!'. While it is important to keep such differences in mind, the general point still stands: Russian is more prone to using the IPV in a pragmatic way and should thus have a higher percentage of IPV imperatives. The very fact that the IPV *davaj* and not the PV *daj* developed into an exhortative particle is, I believe, a testament to that.

3 Corpus study #2: Parallel corpus InterCorp (v9), Czech and Russian

For the second corpus study, I used Alexandr Rosen's parallel corpus InterCorp (v9). I looked for non-negated Czech PV imperatives with an IPV imperative as a Russian equivalent, and vice versa, in both singular and plural. Note that singular and plural are strictly morphological categories here and that both the Czech and the Russian plural covers informal address of a group as well as formal address of individuals or groups. Including Polish with its more complex (pro)nominal system of formal address in this second study was beyond the scope of this paper.

If the pragmatic use of the IPV imperative is in fact much more restricted in Czech, as Benacchio (2010) claims, then there should be more Cz.PV-Ru.IPV correspondences than the other way round. Because translations are aligned by sentence in InterCorp, I went through the results manually to remove any mistakes, e.g. where a Russian IPV imperative just happens to appear in a sentence but is not, in fact, a translation or equivalent of a Czech PV imperative. Figure 1 shows the results for both singular and plural pairings of Cz.PV-Ru.IPV and Cz.IPV-Ru.PV.

As expected, in both singular and plural the pairing Cz.PV-Ru.IPV is more frequent than Cz.IPV-Ru.PV, possibly due to the more widespread use of the pragmatic IPV in Russian. In the following two examples, the Russian IPV can in fact be interpreted pragmatically as "urging":

(7) a. Míšo, prosím tě, prijeď domů. (Cz)
Míša.VOC ask.1SG.PRES you.ACC come-(driving).PV.IMP.SG to-home
b. Miša, požalujsta, priezžaj domoj. (Ru)
Miša please come-(driving).IPV.IMP.SG to-home
'Miša, please come home.' (InterCorp v9)

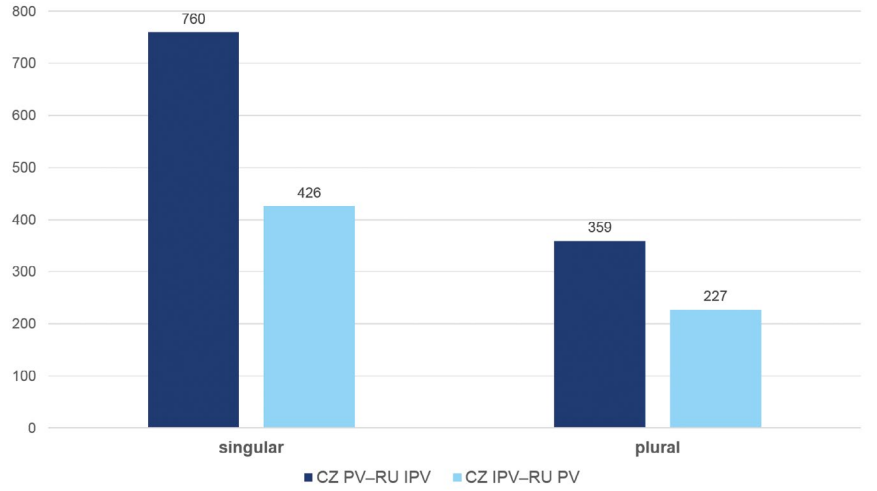


Figure 1: Search results InterCorp (v9), Czech and Russian pairings.

(8) a. Kupte si ho bez řečí. (Cz)
buy.PV.IMP.PL REFL.DAT 3SG.M/N without speeches
b. Pokupajte bez razgovorov (Ru)
buy.IPV.IMP.PL without conversations
'Buy it without talking too long.' (InterCorp v9)

One might also expect that the observed asymmetry between Czech and Russian is due to the fact that Czech also allows PV in iterations, which Russian disprefers. However, iterative examples are in fact very rare in our sample. One example of this is given in (9):

(9) a. Na každé stanici si kup zpátečný_lístek. (Cz)
on every stop REFL.DAT buy.PV.MP.SG return-ticket
b. Prosto na každoj ostanovke pokupaj po (Ru)
obratnomu_biletu.
simply on every stop buy.PV.IMP.SG one-each-of
return-ticket
'Just buy a return ticket on every stop.' (InterCorp v9)

The most frequent Czech PV imperatives translated using a Russian IPV are (singular only): *podívej* 'look' (139), *posaď (se)* 'sit down' (28), *poslechni* 'listen'

(24), *vrať (se)* 'return' (24), *zůstaň* 'remain' (21), *rozděl* 'divide' (19)[2], *sedni (si)* 'sit down' (20), *spusť* 'start, get going' (19), *odpověz* 'answer' (15), *chyť* 'grab' (14).

The most frequent Czech IPV imperatives to be translated with a Russian PV are (again, only singular): *poslouchej* 'listen' (42), *pojď*[3] 'come' (37), *pamatuj* 'remember' (35), *jdi* 'go' (30), *věř* 'believe' (23), *běž* 'run' (21), *drž* 'hold' (19), *mlč* 'be silent' (18), *povídej* 'tell' (17), *snaž (se)* 'try' (13).

One can ask, of course, why there are any Cz.IPV-Ru.PV pairings at all. When we look at the Czech lexemes in question here, we find that a quarter of these cases belong to "partnerless" IPV verbs and do not therefore participate in the aspectual opposition. These are motion verbs[4] like *jít* 'go', *běžet* 'run', but also *vyprávět* 'tell (a story)' and *držet* 'hold'[5]. We can speculate that maybe they would express their imperative in a PV form if they could. Figure 2 is an update of Figure 1, with the added column showing Cz.IPV-Ru.PV pairings with these partnerless verbs removed, which makes the asymmetry even more pronounced.

Regarding Czech IPV *poslouchej* 'listen'; I believe that in these cases listening is seen as an atelic "state of paying attention", hence the IPV. The fact that they correspond to PV Russian *poslušaj* is due to the nature of the Russian prefix *po-*, which can convey a delimitative meaning of 'doing s.th. for some time' (cf. AG-80:365), combining perfectivity and atelicity, whereas Czech does not have this option and thus by necessity uses the IPV to convey atelicity. When combined with a concrete object, the listening becomes telic and the PV becomes the preferred choice in Czech as well.

2 All of these are part of the fixed expression *rozděl a panuj* going back to Latin *divide et impera* (conventionally rendered into English as 'divide and conquer'). This is not a true imperative but rather a name for a certain strategic approach.

3 *pojď* is a second imperative of *jít* 'go' next to *jdi.* The difference is not one of aspect, however: *pojď* is used to mean 'come here!', whereas *jdi* means 'go away!'

4 Note that Czech simplex motion verbs are peculiar in this regard, as the Russian equivalents of these motion verbs do have a PV partner, as do lexically derived motion verbs in Czech, such as *odejít*.PV – *odcházet*.IPV 'go away'. There is some confusion as to the aspectual nature of Czech simplex motion verbs, especially because their preterite is often used like a PV verb might. Since they can, however, be used in progressive contexts, which prohibit the PV, I opt for describing them as IPV, possibly biaspectual in the preterite. This is not of direct import for this study, however, because they still do not partake in a formal aspectual opposition with a partner verb.

5 18 of 19 tokens of *drž* 'hold' are part of the phraseologism *drž hubu!* 'shut up!', possibly a Germanism.

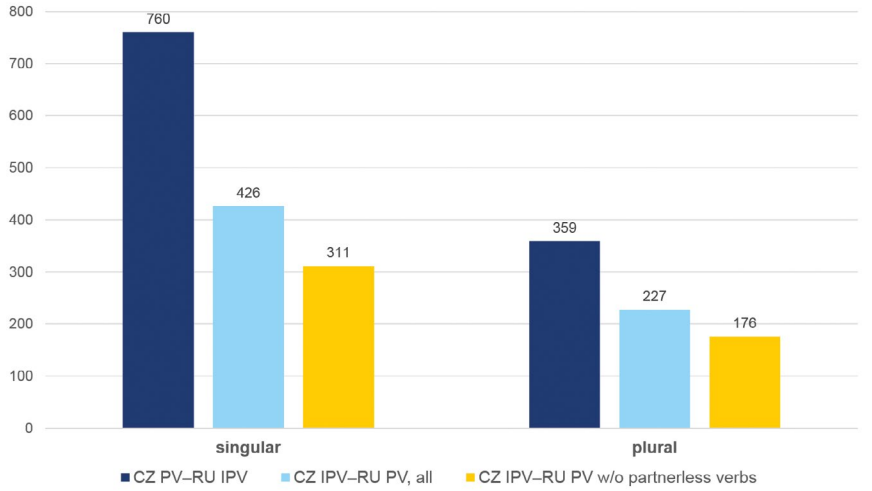


Figure 2: Search results InterCorp (v9), Czech and Russian pairings w/o partnerless verbs.

4 Conclusion

Czech and Russian clearly show an asymmetry in the way they use verbal aspect in the imperative, confirming our prediction based on Benacchio's claims. The first study, which also included Polish, has shown that in most cases the relative frequency of the IPV imperative is higher in Russian than it is in Czech and Polish. In the second study we have seen that Cz.IPV-Ru.PV pairings are less frequent than Cz.PV-Ru.IPV, which points us in the same direction: Russian uses the IPV imperative more often than Czech does. While the immediate context (phraseologisms) and lexical idiosyncrasies (missing partner verbs, language-specific additional meanings of a given verb form) certainly play a role as well, this asymmetry between Czech and Russian is at least partially due to a difference in pragmatics between the two, namely the widespread use of the pragmatically-motivated „polite" or „rude" IPV imperative in Russian.

References

[AG-80]: Švedova, N. Ju. (ed.) 1980. *Russkaja grammatika.* Moskva: Nauka. online at: http://www.rusgram.narod.ru.

Benacchio, Rosanna. 2010. *Vid i kategorija vežlivosti v slavjanskom imperative. Sravnitel'nyj analiz.* München, Berlin: Sagner.

Benko, Vladimír. 2015. *Srovnatelné webové korpusy Aranea.* Ústav Českého národního korpusu FF UK, Praha. http://www.korpus.cz.

Dickey, Stephen M. 2000. *Parameters of Slavic Aspect. A Cognitive Approach.* Stanford: CSLI Publications.

Dübbers, Valentin. 2015. Factors for Aspect Choice in the Different Contexts of Open Iteration in Czech. In Rosanna Benacchio (ed.), *Verbal Aspect: Grammatical Meaning and Context,* 197–210. München: Sagner.

Eckert, Eva. 1984. *A Contrastive Study of Czech and Russian Aspect.* Ann Arbor: University Microfilms International.

[InterCorp v9]: Alexandr Rosen and Martín Vavřín. 2016. *InterCorp verze 9.* Ústav Českého národního korpusu FF UK, Praha. http://www.korpus.cz.

Lehmann, Volkmar. 2008. Pragmatische Quasi-Synonymie: Zur Höflichkeit der russischen Aspekts. In Alicja Nagórko et al. (eds.), *Sprache und Gesellschaft: Festschrift für Wolfgang Gladrow,* 151–155. Frankfurt/Main: Peter Lang.

Lehmann, Volkmar. 2009. Formal-funktionale Theorie des russischen Aspekts. http://subdomain.verb.slav-verb.org/Aspekttheorie.html.

[NKJP]: *Narodowy korpus języka polskiego.* http://www.nkjp.pl.

[NKRJa]: *Nacional'nyj korpus russkogo jazyka.* http://www.ruscorpora.ru.

Padučeva, E.V. ²2010. *Semantičeskie issledovanija. Semantika vremeni i vida v russkom jazyke. Semantika narrativa.* Moskva: Jazyki slavjanskoj kul'tury.

Stunová, Anna. 1993. *A Contrastive Study of Russian and Czech Aspect: Invariance vs. Discourse.* PhD thesis, Universiteit van Amsterdam.

[SYN2015]: Křen, Michal et al. 2015. *SYN2015: reprezentativní korpus psané češtiny.* Ústav Českého národního korpusu FF UK, Praha. http://www.korpus.cz.

von Waldenfels, Ruprecht. 2012. Aspect in the Imperative across Slavic – a Corpus Driven Pilot Study. In Atle Grønn, Anna Pazelskaya (eds.), *The Russian Verb.* Oslo Studies in Language 4(1), 141–154.

Wiemer, Björn. 2001. Aspect choice in non-declarative and modalized utterances as extensions from assertive domains (Lexical semantics, scopes, and categorial distinctions in Russian and Polish.) In Hauke Bartels, Nicole Störmer, Ewa Walusiak (eds.), *Untersuchungen zur Morphologie und Syntax im Slavischen.* Oldenburg: BIS.

Wiemer, Björn. 2008. Zur innerslavischen Variation bei der Aspektwahl und der Gewichtung ihrer Faktoren. In Karl Gutschmidt, Ulrike Jekutsch, Sebastian Kempgen (eds.), *Deutsche Beiträge zum 14. Internationalen Slavistenkongress, Ohrid 2008,* 383–409. München: Sagner.

Björn Hansen, Zrinka Kolaković, Edyta Jurkiewicz-Rohrbacher

Clitic Climbing and Stacked Infinitives in Bosnian, Croatian and Serbian – A Corpus-Driven Study[1]

Abstract Although clitics (CLs) have been very often analysed for Bosnian, Croatian and Serbian (BCS), only few studies approach clitic climbing (CC) in BCS. According to Čamdžić & Hudson (2002) and Aljović (2004), CC out of infinitive complements is obligatory. In the present paper, we focus on constructions with stacked infinitives and address the following research question: "Can pronominal CC appear in the context of stacked infinitives?" Based on material extracted from three web corpora {bs, hr, sr}WaC, we conclude that pronominal CC does not always occur in the case of stacked infinitives in all three languages examined. We identify the following constraints: 1. CLs in the same case but depending on two different verbs block CC. 2. Reflexivity of the infinitive embedding further infinitives seems to be involved in the blocking of CC.

Keywords Clitic climbing, stacked infinitives, web corpora, Bosnian, Croatian, Serbian

1 Introduction

The syntax of clitics in Bosnian, Croatian and Serbian, by some authors called Serbo-Croatian (BCS), has been the target of intense theoretical research. The placement of clitics (CL) is usually associated with the left edge of the sentence, the so-called 'second position'. Most works on CL in Bosnian, Croatian and Serbian address the nature of this second position effect, mainly within formal

1 This study was carried out within the research project 'Microvariation of the Pronominal and Auxiliary Clitics in Bosnian, Croatian and Serbian. Empirical Studies of Spoken Languages, Dialects and Heritage Languages' funded by the Deutsche Forschungsgemeinschaft (HA 2659/6-1, 2015-2018). We are grateful to the anonymous reviewer for helpful comments on an earlier version of the paper.

theoretical frameworks (primacy of syntactic vs prosodic processes, for an overview see Bošković 2004, Browne 2003, 2004, 2014, Franks & King 2000, Franks 2010). Descriptively speaking, CLITIC CLIMBING (CC) refers to sentence structures in which "the clitic is associated with a verb complex in a subordinate clause but is actually pronounced in construction with a higher predicate (for instance, the matrix verb which selects that subordinate clause), even though it may have no obvious semantic or syntactic connection to that verb" (Spencer & Luís 2012: 162). An example of CC out of an infinitival complement is given in (1) where the clitical pronoun *ga* has to move from the infinitival into the matrix clause:

(1) Marija **ga**$_2$ mora$_1$ vidjeti$_2$.
Marija him.ACC must.3PRS see.INF

(1') *Marija mora$_1$ vidjeti$_2$ **ga**$_2$.
'Marija must see him.' Aljović (2004)

Čamdžić & Hudson (2002: 326) argue that in BCS CC "[…] is obligatory when the complement is an infinitival form and marginally possible when the complement is a *da* clause". Two years later Aljović (2004) claims the same: in the case of restructuring verbs, CC out of infinitive complements "is not an option but a necessity". However, they do not provide any empirical evidence. A further work dealing with CC is Stjepanović (2004) but her focus is on *da*-constructions where CC is claimed to be optional. There are, actually, no empirical studies specifically dealing with CC in BCS based on natural data. The syntactic conditions of CC are thoroughly described only for Czech by Junghanns (2002), Dotlačil (2004), Rezac (2005) and Hana (2007) who propose several constraints on CC in this West Slavonic language. As we assume that the word order behaviour of clitics is based on syntactic constraints, we shall refrain from conjecturing about restrictions imposed by allegedly prosodic features.

2 Research question

The few existing studies which mention CC in BCS focus on the structure 'complement taking predicate + infinitive' as in (1), none of them, however, deals with what we call STACKED INFINITIVE CONSTRUCTIONS, i.e. complement taking predicates (CTP) showing multiple embedding of two or more infinitives, as in example (2):

(2) Pokušavao$_1$ je prestati$_2$ pušiti$_3$, (…)
try.PTCP.SG.M be.3SG stop.INF smoke.INF
'He tried to quit smoking, (…)' (hrWaC v2.2)

We believe that precisely stacked infinitives are an ideal test case for constraints on CC because they contain all types of combinations of CL and therefore allow to identify possible contexts of blocked CC (on Czech see Hana 2007: 122–132). A further reason to restrict the search to stacked infinitives is a methodological one. Since the structure CTP + one infinitive is rather frequent, we would have been forced to work with samples that would not have enabled us to detect possible constraints, since in the samples frequently occurring raising CTP-like e.g. modal or phrasal verbs would have predominated.

In the following, we are going to test Čamdžić & Hudson's (2002: 326) and Aljović's (2004) claim that CC is obligatory in infinitival complements. Our research question is:

> "Is CC obligatory in the context of stacked infinitives (embedding of two or more infinitives)?; i.e. can stacking of infinitives block CC?"

Our study is corpus-driven; we will present the actually attested constructions and their frequencies.

3 Data extraction & methodology

We extract the data from three massive, morphosyntactically tagged web corpora: bsWaC v1.2, hrWaC v2.2 and srWaC v1.2 (Ljubešić & Klubička 2014). We look for CLs in three different positions in the context of infinitive stacking (we allowed 2 to 4 infinitives in a row). The following examples (3), (4) and (5) illustrate the possible positions of the clitics:

			CTP	Infinitive	Infinitive	CL
(3)	I	vi	možete$_1$	pomoći$_2$	zaustaviti$_3$	**ga**$_3$ (…)
	and	you.NOM	can.2PRS	help.INF	stop.INF	him.ACC
	'You can also help to stop him (…)'				(bsWaC v1.2)	

	CTP	CL		Infinitive	Infinitive
(4)	Morate$_1$	**ih**$_3$	samo	znati$_2$	prepoznati$_3$.
	must.2PRS	them.ACC	only	know.INF	recognize.INF
	'You just have to know how to recognize them.' (srWaC v1.2)				

		CL	CTP	Infinitive	Infinitive
(5)	Ona	**nas**$_3$	mora$_1$	naučiti$_2$	kontrolirati$_3$.
	she.NOM	us.ACC	must.3PRS	learn.INF	control.INF
	'She has to learn how to control us.' (bsWaC v1.2)				

In example (3), the pronominal CL *ga* remains *in situ*, following its infinitival governor *zaustaviti*. In (4), however, the pronominal clitic *ih*, which is a complement of the infinitive *prepoznati*, climbed into the matrix clause and follows the higher CTP *morati*. A structurally similar situation is found in (5), where the pronominal clitic *nas*, which is a complement of the infinitive *kontrolirati* moved to the matrix clause and precedes the higher CTP *morati*. Both (4) and (5) are perfect examples of CC. Nevertheless, they differ in respect to the word order. Therefore, our queries accounted for both above described word order patterns[2]. Here is example of the query CTP + INF (2,4) + CL[3]:

```
[!(word="(me)|([mj]u)|(joj)|(i[hm])|(ga)|([nv]as)")|
(word="je"&tag!="V.*")|(word="[mt]i"&tag!="(Pp[12]-[sp]
n|Pd-mpn)")|(word="te"&tag!="(Pd-
[fm][sp][nga])|(Cc)")]{1,4}[tag="Vm.*"&lemma!="ht-
j?eti"&word!="nemoj[mt][eo]"][!(tag="C.*"|lemma="\Z"|
tag="P[iq].*"|(tag="(V.*y)|(Va[aefpmn].*)|(Vc[ef-
pmn].*)|(Vm.*)"&word!="ćeš")|tag="Rr"|word=".*\..*"|-
lemma="što")]{0,4}[tag="V.n"]{2,4}[!(tag="C.*"|-
lemma="\Z"|tag="P[iq].*"|(tag="(V.*y)|(Va
[aefpmn].*)|(Vc[efpmn].*)|(Vm.*)"&word!="ćeš")|tag="R-
r"|word=".*\..*"|lemma="što")]{0,4}[(word="(me)|([mj]
u)|(joj)|(i[hm])|(ga)|([nv]as)")|(word="-
je"&tag!="V.*")|(word="[mt]i"&tag!="(Pp[12]-[sp]n|Pd-mp-
n)")|(word="te"&tag!="(Pd-
[fm][sp][nga])|(Cc)")]{1,4}[!(tag="R-
r"|(tag="(V.*y)|(Va[aefpmn].*)|(Vc[efpmn].*)|(Vm.*)"&word!
="ćeš")|word=".*\..*"|lemma="što")]{0,4}within<s/>
```

2 We are aware of the fact that the reverse order infinitive complements-CTP is possible in BCS, but we did not take it into account, because it represents information about structurally marked word order. Additionally, infinitive + infinitive + CTP poses difficulties for corpora, where the sentence clause border is not annotated. Hence, the precision of the queries in question would be very low.

3 Index of morphosyntactic descriptions MSD at http://nl.ijs.si/ME/V5/msd/html/msd-hr.html#msd.msds-hr.

In the query, we excluded all forms of the lemma *htjeti* ('will', 'want') since the corpus annotation does not offer disambiguation of its function as an auxiliary verb, which in combination with the infinitive forms the future tense, or as modal verb. Furthermore, we excluded the forms *nemoj*, *nemojmo* and *nemojte*, which in combination with the infinitive express prohibitive in BCS.

In order to obtain most occurrences of the constructions, we could not restrict the query only to the core elements of the construction (CTP, Infinitive stack, CL), but we allowed empty positions, so elements such as clitics governed by CTP could appear. Nevertheless, we excluded from empty positions most elements marking the sentence clause, such as conjunctions, other main verbs, and punctuation signs.

The resulting recall required manual processing, also due to errors in tagging. Since hrWaC v2.2 is two and a half times bigger than srWaC v1.2 and five times bigger than bsWaC v1.2 the query returned proportionally higher results, which are almost impossible to process manually. Therefore, for hrWaC v2.2 we generated three samples via NoSketch Engine (function "Sample") which comprise a quarter of the originally retrieved hits.

Apart from empty positions which decreased the recall, some duplicates and hits which were linked only to CC out of the first infinitive, as in the example given in (6), had to be excluded manually.

(6) (…) možemo$_1$ **im**$_2$ pomoći$_2$ popraviti$_3$ ponašanje (…)
can.1PRS them.DAT help.INF correct.INF behaviour.ACC
'(…) we can help them to correct their behaviour (…)' (bsWaC 1.2)

The reason for that is the fact that in accordance with our research question we focus on CL depending on the second infinitive, as is it only in that case that stacked infinitives may or may not block CC. The sentences in which two clitics appeared, one as a complement of the first infinitive and the other as a complement of the second (or in rare cases of the third) infinitive were taken into consideration, see the example in (7):

(7) (…) možete$_1$ **si**$_2$ dozvoliti$_2$ uskratiti$_3$ **mi**$_3$ sve (…)
can.2PRS REFL.DAT allow.INF curtail.INF me.DAT everything
'(…) you can allow yourself to curtail everything from me (…)' (hrWaC v2.2)

In those cases, our focus was on the clitic which is a complement of the second infinitive (here *uskratiti*) and, of course, on the relationship between two clitics, by which we mean the formation of a clitic cluster or clitic split as in the case of *si* and *mi* in example (7).

Although our queries allowed a maximum of four embedded infinitives, we found only three examples with three infinitives (see one of the examples in (8)) and no example of a bigger stack.

(8) (…) samo **se**$_3$ ne smijem$_1$ zaboraviti$_2$ sjetiti$_3$
only REFL NEG must.1PRS forget.INF remember.INF
reći$_4$ **im**$_4$ (…)
tell.INF them.DAT (hrWaC v2.2)
'(…) I only must not forget to remember to tell them (…)'

A corpus-driven study may help to determine factors which are responsible for CC or the lack of CC respectively, but it requires an additional manual annotation of samples. In the present study, our annotation scheme contains the language variety, the word order behaviour of CL, grammatical features of the CL and basic syntactic properties of the predicates the CL depends on (raising vs control).

4 Results & discussion: clitic climbing and stacked infinitives

Our results give a clear answer to the research question. As can be seen in Figure 1, which presents the final distribution of the target constructions across each corpus, stacked infinitives as such do not prevent CLs from climbing into the matrix clause. We find both examples with CC (83,44–86,12 %) and without CC (13,88–16,56 %).

We have not found significant, language-specific differences in the distributions of the constructions with CC and without CC (χ^2 test, p-value 0.51). The low overall recall in srWaC v1.2 can be explained by the fact that especially in Serbian the infinitive competes with the semifinite *da*-construction, as in (9).

(9) (…) stvarno moram$_1$ da počnem$_2$ da učim$_3$ (…)
really must.1PRS COMP start.1PRS COMP learn.1PRS
'(…) I really have to start to study (…)' (srWaC v1.2)

Regarding our results, it is interesting to point out that even in those rare cases with three infinitives. The CL of the last infinitival complement could climb over three CTPs into the matrix clause, as shown in (10): *držati ga* ('to hold him')

(10) (…) i u svakome trenutku **ga**$_4$ možemo$_1$
and in any moment him.ACC can.1PRS
odlučiti$_2$ prestati$_3$ držati$_4$ (…)
decide.INF stop.INF hold.INF (hrWaC v2.2)
'(…) and in any moment, we can decide to stop holding him (…)'

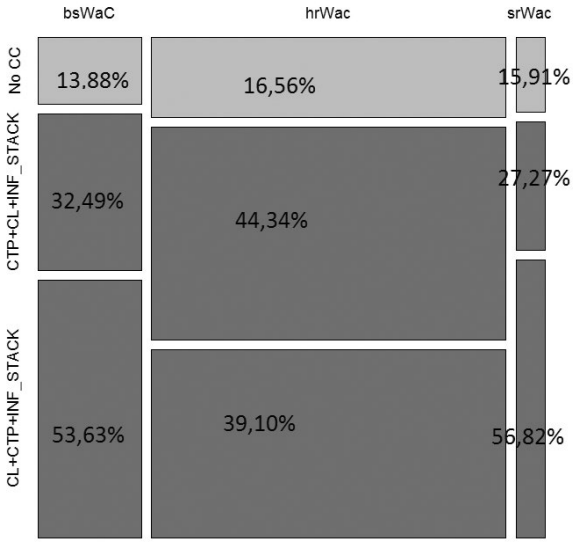


Figure 1: CC and stacked infinitives in {bs, hr, sr}WaC. (Σ of all examples 1492 = 317 (bsWaC v1.2) + 1087 (hrWaC v2.2) + 88 (srWaC v1.2))

5 Conclusion & further perspectives

To conclude, our corpus-driven study based on data from three web corpora has shown the range and frequency of word order patterns of CL in constructions with stacked infinitives in BCS. We have found that:

i. Clitics can climb within stacked infinitives.
ii. In stacked infinitive constructions, CC is found in around 83,44–86,12 % and the lack of CC in 13,88–16,56 % of all cases.
iii. There are no significant, language-specific differences in the distributions of the researched constructions.

Coming back to our research question from Section 2, we can draw the conclusion that CC in BCS is not always obligatory (contra Čamdžić & Hudson 2002: 326). This might be explained in two ways: first, CC *per se* is facultative or, second, CC is obligatory but subject to constraints.

Following the latter assumption, our corpus-driven study allows formulation of a few hypotheses concerning possible constraints on CC:

i. We found some evidence for 'Same case different governors constraint': CC might be blocked if two CL depending on two different CTPs have the same case as in ex. (7) where two clitics in Dative are split (*si, mi*). It is worth pointing out that this constraint may be a subtype of 'object control case constraint' (see Dotlačil 2004 and Rezac 2005 for more details).

ii. Reflexivity of the infinitive embedding further infinitives seems to play a crucial role in blocking clitic climbing (Odds Ratio test with 95% confidence level yields 502.8000, $p<0.0001$) as in ex. (7) and (8)[4].
iii. We have also found a significant relation between the syntactic type of the infinitive governing further infinitives (Chi-square test 95.78, $p<0.0001$), but with medium size effect (Cramer's $V=0.2535$). CC from infinitive stacks governed by object-control infinitive (as the predicate *pomoći* 'to help' in ex. (3)) or by subject-control infinitive is more restricted than from raising. Our findings from (ii) help explain this fact: raising verbs are never reflexive, while every sixth subject-control and every eighth object-control verb in our data set is reflexive.

More findings could be obtained by extending the annotation schema. In the future, we intend to explore whether grammatical or lexical properties of the CL themselves influence CC, and how CL interacts with CL governed by other infinitives and CTP. This will allow a clearer picture of the nature of CC.

We have to be aware however, of the fact that the patterns of actual language usage described in this paper do not directly reflect constraints in a proper sense of the word. A corpus study can only provide first clues for possible constraints on CC. As not all combinations of CTPs and CL could be found in the corpora we envisage the triangulation of methods; i.e. we plan to carry out systematic experiments comprising acceptability judgements with a larger number of native speakers. As argued by Diesing, Filipović Đurđević & Zec (2009), the study of the syntax of clitics demands the combination of corpus and experimental data.

6 References

Aljović, Nadira. 2004. Cliticization Domains: Clitic Climbing in Romance and in Serbo-Croatian. In Olivier Crouzet, Hamida Demirdache and Sophie Wauquier-Gravelines (eds.), *Proceedings of JEL'2004 Domain(e)s*, 169–175. Université de Nantes.

Aljović, Nadira. 2005. On clitic climbing in Bosnian/Croatian/Serbian. In Nedžad Leko (ed.), *Lingvistički vidici* 34:(05), 58-84. Sarajevo: Forum Bosnae.

4 One of the anonymous reviewers proposed the stricter formulation "reflexivity blocks CC", we, however, would like to keep it in this way since in this first phase of our research we did not distinguish between different types of reflexive CLs. Lešnerová & Malink's (2008) study conducted on Czech suggest that different reflexives indeed behave differently in respect to CC. This may be the case in BCS as well and we plan to investigate it in more depth.

Bošković, Željko. 2004. Clitic placement in South Slavic. *Journal of Slavic Linguistics* 12: (1), 37–90.

Browne, Wayles. 2003. Razlike u redu riječi u zavisnoj rečenici. *Wiener Slawistischer Almanach*. Sonderband 57, 45–52.

Browne, Wayles. 2004. Serbo-Croatian Enclitics for English-Speaking Learners. *Journal of Slavic Linguistics* 12: (1), 249–283 (Reprint from 1975).

Browne, Wayles. 2014. Groups of Clitics in West and South Slavic Languages. In Elżbieta Kaczmarska and Motoki Nomachi (eds.), *Slavic and German in Contact: Studies from Areal and Contrastive Linguistics*. Slavic Eurasian Studies 26, 81–96. Hokkaido: Slavic Research Center.

Čamdžić, Amela; Hudson, Richard. 2002. *Clitics in Serbo-Croat-Bosnian*. In John Harris et al. (eds.), *UCL Working Papers in Linguistics 14*, 321–353. London: Department of Phonetics and Linguistics University College London. http://www.phon.ucl.ac.uk/home/PUB/WPL/02papers/camdzic_hudson.pdf. (23.06.2017).

Diesing, Molly; Filipović Đurđević, Dušica and Zec, Draga. 2009. Clitic placement in Serbian: Corpus and experimental evidence. In Susanne Winkler and Sam Featherston (eds.), *The Fruits of Empirical Linguistics II: Product*, 59–73. Berlin/New York: Mouton de Gruyter.

Dotlačil, Jakub. 2004. *The syntax of infinitives in Czech*. Master's Thesis. University in Tromsø. http://jakubdotlacil.com/thesis.pdf (03.05.2017).

Franks, Steven. 2010. Clitics in Slavic. *Contemporary Issues in Slavic Linguistics* 10, 1–157.

Franks, Steven; King, Tracy Holloway. 2000. *A Handbook of Slavic clitics*. New York/Oxford: Oxford University Press.

Hana, Jirka. 2007. *Czech Clitics in Higher Order Grammar*. PhD Thesis. The Ohio State University.

Junghanns, Uwe. 2002. Clitic climbing im Tschechischen. *Linguistische Arbeitsberichte* 80, 57–90.

Lešnerová, Šárka; Malink, Marko. 2008. Clitic Climbing and Theta-Roles in Upper Sorbian and Czech. In Gerhild Zybatow et al. (eds.), *Formal Description of Slavic Languages: FDSL 5*, 396-407. Frankfurt am Main: Peter Lang. http://philosophy.uchicago.edu/faculty/files/malink/14%20Clitic%20Climbing%202008.pdf. (16.05.2017).

Ljubešić, Nikola; Klubička, Filip. 2014. {bs,hr,sr} WaC – Web corpora of Bosnian, Croatian and Serbian. In Felix Bildhauer and Roland Schäfer (eds.), *Proceedings of the 9th Web as Corpus Workshop* (WaC-9), 29–35. Gothenburg. Sweden.

Rezac, Milan. 2005. The syntax of clitic climbing in Czech. In Lorie Heggie and Francisco Ordóñez (eds.), *Clitics and affix combinations. Theoretical perspectives*, 103–140. Amsterdam: Benjamins.

http://minimalism.linguistics.arizona.edu/AMSA/PDF/AMSA-202-0602.pdf. (02.05.2017).

Spencer, Andrew; Luís, Ana R. 2012. *Clitics. An Introduction.* Cambridge: Cambridge University Press.

Stjepanović, Sandra. 2004. Clitic Climbing and Restructuring with "Finite Clause" and Infinitive Complements. *Journal of Slavic Linguistics* 12: (1), 173–212.

II. Methodology and Application

Alexandr Rosen

Coping with Unruly Language: Non-Standard Usage in a Corpus

Abstract A language as used in real situations may differ substantially from its standard form. Before the entire range of NLP methods and tools can be applied to non-canonical variants of a language, appropriate categories for the analysis of deviant forms and constructions are needed, together with texts annotated by these categories. A discussion of non-standard language is followed by two case studies. The first study proposes a taxonomy of morphosyntactic categories as an attempt to analyze non-standard forms in non-native learners' Czech. The second study focuses on the role of a rule-based grammar and lexicon as tools for the detection and diagnostics of non-standard words and constructions in the process of building and using a parsebank.

Keywords Non-standard language, Czech, learner corpus, parsebank, treebank, constrain-based grammar, valency, HPSG

1 Introduction

In most cases, corpus annotation is not explicit about the canonicity of language use, although exceptions exist in specialized corpora or in specific cases in mainstream corpora (individual word forms – colloquial, dialectal or non-words). Non-standard usage defies general rules of grammar – it may involve performance errors, creative coinages, emerging phenomena. We start with the assumption that the text in a corpus and its linguistic annotation is where the two Saussurean faces of a single coin converge: the empirical evidence (language use, parole, performance, corpus) and the theory (language as a system, langue, competence, grammar). The annotation is also where multiple levels of analysis and linguistic theories may meet. An annotation scheme defined in terms of appropriate categories or even as a formal grammar can help to identify the difference between the regular and irregular, between the language as a system and its use.

It is often the case that instances of language use – in writing or speech of native and non-native speakers alike – do not comply with a norm or conventional pattern. The need to process non-standard language is growing, especially due to its ever more prominent presence in social media and the stepwise erosion of the role of language variants as social symbols or appropriate vehicles of communication, but also due to the increasing share of non-native speakers in many communities. The latter has additional consequences on the didactic front, represented mainly by the need to develop better methodologies suited to the non-native learner of a specific language.

Interestingly, linguistic variation impedes human communication only to a limited extent. Language users are able to recover meaning from idiosyncrasies on any level of the linguistic system and even recognize signals conveyed by the deviations to make guesses about the speaker's background or intention. On the other hand, standard NLP tools are usually much less adaptive and efficient when applied to non-standard language. Rule-based models, apparently vulnerable to any unexpected phenomena due to their dependence on (under-developed) conceptual categories and frameworks, are at a clear disadvantage. Stochastic models, generally more robust, seem to be in a better position. Possible strategies include applying a model trained on standard language, annotating more data, normalizing test data, deliberately corrupting training data, or adapting models to different domains. Eisenstein (2013) stresses the importance of a suitable match between the model and the domain of the text, while Plank (2016) points out that rather than to domains, the tools should be adapted to text varieties in a multi-dimensional space of factors such as dialect, topic, genre, gender, age, etc. Anyway, at least for rule-based or supervised models we lack suitable concepts and frameworks even distantly comparable to those for standard language. This leads us back to the issue of a suitable taxonomy and markup of unexpected phenomena – one of the topics of this paper (see section 3).

A rationalist approach to modeling non-standard language varieties has an important role not only in the design of categories suited for the analysis of non-standard forms and structures. Rather than being a random collection of unrelated phenomena, each variety represents a system, with rules and principles partially shared with other varieties, standard or non-standard. Deviations from the standard often represent regularly occurring patterns, such as spelling errors due to attraction in subject-predicate agreement.[1] There are many other regular phenomena which occur in the process of acquisition of non-native

1 A 100M corpus of Czech (SYN2010, see http://korpus.cz) includes 47 instances of short distance subject-predicate agreement patterns including spelling errors in masculine animate past tense forms, where the *-ly* ending is used instead of the correct homophonous *-li* ending (Dotlačil 2016).

language, some of them universal or specific to the target language, some of them due to the influence of the native or some other language already known to the learner. These deviations reveal facts about the speaker, her target and native language and can be used in methods and tools identifying the speaker and her background. Discovery of these rules and principles has practical benefits for foreign language teaching, forensic linguistics, the identification of the author's first language or the processing of non-standard language in general.[2]

A general discussion of issues related to non-standard language (section 2) is followed by two case studies. The first study (section 3) presents a taxonomy of learner language phenomena as an attempt to analyze non-standard forms produced by non-native speakers of Czech. The second study (section 4) focuses on the role of a rule-based grammar and lexicon as tools for the detection and diagnostics of non-standard words and constructions in the process of building and using a parsebank.

2 Non-standard language and its types

What counts as non-standard language? According to Bezuidenhout (2006), non-standard use of a language is one that "flouts a linguistic convention or that is an uncommon or novel use." The standard, conventional use is based on an explicit or implicit agreement among members of a linguistic community about the appropriate form of the language, given a specific situation.

This definition is problematic – it may not include some common language varieties that are quite far from the assumption about a standard, both in traditional linguistics or in NLP, such as Twitter messages. It might be useful to position specific varieties within a space of oppositions: the prescriptive or literary norm in contrast to colloquial, dialectal, 'uneducated' or archaic use; the language as a system (*langue*, the idealized linguistic competence) in contrast to the real use of language (*parole*, linguistic performance); written in contrast to spoken varieties; native in contrast to non-native language; the language of a child in contrast to the language of an adult native speaker; the language of people without language disorders in contrast to those with such handicaps; and also expectations of the grammar writer in contrast to anything else. Then we could delineate our notion of non-standard language to include varieties: (i) as used beyond the community of native speakers, (ii) of non-literary language (iii) of spoken language, and (iv) including deviations due to the specifics of language production, i.e. performance errors of all sorts.

2 E.g. typing assistants could offer an option to handle colloquial forms.

On the other hand, Hirschmann et al. (2007) define 'non-canonical' utterances in learner texts as:

> "[...] structures that cannot be described or generated by a given linguistic framework – canonicity can only be defined with respect to that framework. A structure may be non-canonical because it is ungrammatical, or it may be non-canonical because the given framework is not able to analyze it. For annotation purposes the reason for non-canonicity does not matter but for the interpretation of the non-canonical structures, it does. Most non-canonical structures in a learner corpus can be interpreted as errors [...] whereas many non-canonical structures in a corpus of spoken language or computer-mediated communication may be considered interesting features of those varieties."

This 'technical' view of what counts as non-standard language is more suitable to the tasks of annotating Czech as a foreign language and analyzing non-standard linguistic phenomena in a parsebank of Czech. After all, as Hirschmann et al. (2007) note, even if the interpretation of non-canonical structures differs for non-native and native speakers, many issues related to their appropriate annotation or analysis are shared.

Non-standard language can be detected, diagnosed and annotated by NLP methods in various ways (Meurers 2013; Meurers and Dickinson 2017). Tools developed for standard language and trained on standard or non-standard language can be applied (Ramasamy et al. 2015), texts can be manually annotated to build more task-specific models (Aharodnik et al. 2013), hand-crafted rules targeting relevant varieties can be used. It seems that designing an annotation scheme specific to non-standard language to build such a model brings better results (Berzak et al. 2016) than efforts to shoehorn existing annotation schemes to fit learner data (Cahill 2015). These results point to the need of "non-canonical categories for non-canonical data" (Dickinson and Ragheb 2015). Such categories are not part of common linguistic wisdom. It is not clear how to design a layered taxonomy of errors, an intelligibility metrics or a specification of the influence of other languages. The following section includes a proposal for a taxonomy of some phenomena of non-native Czech.

3 Designing categories for Czech as a foreign language

With the advance of learner corpora, the language produced by non-native speakers has been analyzed from perspectives familiar to corpus linguists but not so common in the field of language acquisition: learner texts are annotated by morphological and syntactic categories and structures, surveyed by statistical tools, and used to build stochastic models. Additional annotation, specific to learner language, has been used to capture non-standard phenomena: deviant forms and structures are assigned *target hypotheses* (corrections) and/or error types. So far, there are no standard solutions to these tasks.[3] Principles of emendation, error taxonomies and the shape of annotation schemes differ between projects, reflecting different answers to questions such as: What aspects of learner languages should be annotated? To what extent should the error taxonomy reflect standard linguistic categories and levels? Should multiple hypotheses be allowed, both in correction and error annotation? Is there any alternative to error annotation linked to a specific target hypothesis or can learner texts be analyzed and annotated as *interlanguage*, a language sui generis, approximating the target language in the process of language acquisition, to some extent independently of the target language?

A common strategy is to base the annotation on the concepts of native speakers' grammar, marking up deviations from the standard language in terms of errors in spelling, morphology, syntax, lexical choice, phraseology or register. However, some of the questions must be answered anyway: a nominal form, supposedly an object argument, marked by an incorrect morphological case, could be an error in spelling, morphology or syntax. An annotation scheme may insist on a single choice among these options or allow for their simultaneous specification as disjunctive hypotheses. Forms that do not match any existing word of the standard language (non-words, out-of-lexicon forms) present additional issues.

One possible starting point is a taxonomy of word classes based on a consistent partitioning along the morphological, syntactic and semantic criteria. These criteria are used as a mix in the definition of the standard sets of 8–10 word classes. For some of them, the three criteria yield the same result, but other classes are heterogeneous. A relative pronoun, defined by its semantic property of referentiality to an antecedent, may have an adjectival declension pattern as its morphological property, but it can be used in its syntactic role in a nominal

3 For examples of some tagsets used to annotate learner language see, e.g., http://merlin-platform.eu or https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko.

position.[4] The class of Czech second position clitics consists of auxiliaries, weak pronouns or particles. Auxiliaries, prepositions and reflexive particles may be seen paradigmatically as parts of analytical paradigms in periphrastic verb forms, nouns in "prepositional cases", inherently reflexive verbs, while the rules of syntax treat the independent functional morphemes as individual syntactic words to make sure that they obey constraints on ordering, agreement or government. Thus, morphology, syntax and semantics take different perspectives, calling for a cross-classification of linguistic units at least along the three dimensions of morphology, syntax and semantics. It has been noted before (Díaz-Negrillo et al. 2010) that a cross-classifying scheme can be applied to texts produced by non-native learners. For English, the use of an adjective in an adverbial position can be analyzed as a mismatch between adverb as the syntactically appropriate category and adjective as the lexical category of the form used by the author of the text. A parallel Czech example is shown in (1), where the adjectival form *krásný* 'beautiful' is used instead of the standard adverbial form *krásně* 'beautifully'. The word can be annotated as a morphological adjective and syntactic adverb.

(1) Whitney Houston zpívala **krásný** → krásně
Whitney Houston sang beautiful → beautifully
'Whitney Houston sang beautifully.'

However, a morphologically rich interlanguage often deviates not just in the use of word classes but also in morphology. In (2), *táta* 'daddy' is nominative, but as the object of *viděl* 'saw' it should be accusative, which could be represented in the cross-classifying taxonomy as a mismatch between morphology and syntax in the category of case. A parallel example in English would be (3)[5] or, with a mismatch in number (4).

(2) Lucka viděla **táta** → tátu
Lucy.nom saw daddy.NOM → daddy.ACC
'Lucy saw her dad.'

(3) I must play with **he**.NOM → him.ACC

(4) The first year **have**.PL → has.SG been wonderful.

4 For a more detailed description of the proposed taxonomy of word classes see Rosen (2014).

5 The example is taken from Dickinson and Ragheb (2015).

In (5), the aspect of the content verb *napsat* 'to write' is perfective, while the auxiliary verb *bude* can only form an analytical future tense with an imperfective form. A perfective verb is used in its present form to express future meaning, as in (6).

(5) Eva **bude** **napsat** dopis
Eva will write.PFV letter
'Eva will write a letter.' (intended)

(6) Eva napíše dopis
Eva writes.PFV letter
'Eva will write a letter.'

Although the cross-classification idea can be applied to the analysis of all the above examples as mismatches between morphology and syntax, it does not seem to be the most intuitive solution. The annotation of (3) is agnostic about the fact that *he* is in a wrong case after all, a fact that should probably be avoided in the annotation of interlanguage, but which seems to be intuitive and important anyway. The form is only nominative rather than both nominative and accusative. While nominative is the morphological category, the missing syntactic interpretation is that of an object, a category specific to the layer of syntax.

The original proposal of Díaz-Negrillo et al. (2010) is concerned with English learner texts, assuming only standard POS labels at three layers: distribution (syntax), morphology and lexical stems. In standard language, the evidence from the three levels converges on a single POS. Mismatches indicate an error: stem vs. distribution (*they are very kind and* ***friendship***), stem vs. morphology (*television, radio are very* ***subjectives***), distribution vs. morphology (*the first year* ***have*** *been wonderful*). All these types are attested in Czech, but due to a wide range of phenomena related to morphonology and morphology, bare POS and mismatches of this type are not sufficient.

Our proposal combines error annotation with "linguistic" annotation of the original and the corrected version of the text, using standard categories such as domain-specific word class and other morphosyntactic properties as far as possible. Linguistic annotation of the original text may thus result in some forms labelled as unknown. Error annotation is based on the relation between the original and the corrected form, and on the relation between their analyses. An error is analyzed from three perspectives: (i) domain (see below), (ii) register (style), which is used as the benchmark to determine the error status, and (iii) location within the form, specified in terms of character positions and – if possible – in terms of a morpheme, such as stem, prefix, derivational suffix or inflectional ending. We propose five domains: spelling, morphonology, morphology, syntax and lexicon. Errors in each of the domains can be specified in more detail.

Spelling errors include word boundaries, punctuation, missing or incorrect capitalization (***mannheim*** →*Mannheim*), confusion of the homophonous vowels *i* and *y* (***lingvistyka*** →*lingvistika*), absence of graphemes such as *ě*, expressing palatalization of a preceding consonant (***ďeti*** → *děti*) or *j* followed by *e* as phonemes (***vjec*** → *věc*), and other issues connected with the use of diacritics.

Morphonology includes problems in palatalization, epenthesis or other processes, such as redundant presence or wrong absence of a vowel in some inflectional paradigms (***pesa*** → *psa* 'dog.ACC.SG' from *pes* 'dog.NOM.SG'; ***sestr*** → *sester* 'sister.GEN.PL' from *sestra* 'sister.NOM.SG'), incorrect presence or absence of vocalized versions of prepositions (***v*** *Vietnamu* → *ve Vietnamu* 'in Vietnam'), or confusion of voiced and devoiced consonants (***sůstala*** → *zůstala* 'stayed'). Given a target hypothesis, most errors in spelling and morphonology can be diagnosed automatically.[6]

Morphology includes paradigmatic errors related to inflectional patterns, including both non-words (*na* ***Erasmuse*** → *Erasmu* 'on the Erasmus'; ***studovám*** → *studuju* 'I study') and existing forms of the given word, inappropriate in the given context. If the original word exists, the error can be morphological or syntactic: *viděla* ***táta*** → *viděla tátu* '[she] saw [her] dad' (2).

Syntax covers syntagmatic issues: word order and incorrect use of word forms in a given context, including improper expression of valency, agreement, quantification etc.

Lexical errors typically concern the use of a semantically or syntactically inappropriate lexeme or even category such as verbal aspect, missing reflexive particle in inherently reflexive verbs, or an issue in phraseology.

It is often difficult to decide about the domain, i.e. about the cause of a specific deviation – is the issue in (2) an error in spelling, morphology or syntax? One possible strategy is to apply a rule selecting a single option. In the manual annotation of the *CzeSL* corpus (Rosen et al. 2014), the rule was to specify the deviation in a domain where the analysis requires a more sophisticated judgment, e.g. morphology or syntax in preference to spelling. An alternative strategy is to specify the deviation in parallel in all relevant domains. This solution leaves the decision open for additional analysis and fits well in the concept of cross-classification.

The combined error and linguistic annotation can be used to tag the corpus and to specify types located within a hierarchy of learner language phenomena. The error annotation together with the two poles of linguistic annotation – one for the ill-formed and one for the corrected word – represent a pattern. For a

6 See Jelínek et al. (2012) for a list of "formal errors": missing or redundant character, character metathesis, etc., which can often be interpreted in linguistic terms.

simple case such as (2), the pattern is shown in Table 1.[7] A taxonomy of such patterns can be built, and references to more or less abstract patterns can be used as tags. A more abstract pattern in Table 2 represents all cases where a nominative form is used instead of an accusative form.

Table 1: The pattern for *táta* in (2) (*Lucka viděla* ***táta*** → *tátu* 'Lucy saw her Dad').

		error annotation	linguistic annotation	
			original	target
location		inflectional suffix	–	–
register		standard	–	–
domain	spelling	character replacement	a	u
	morphology	case	nominative	accusative
	syntax	valency	object of *viděla*	object of *viděla*

Table 2: The abstract pattern for a form which is nominative instead of accusative.

		error annotation	linguistic annotation	
			original	target
location		inflectional suffix	–	–
register		standard	–	–
domain	morphology	case	nominative	accusative

A different type of error is shown in (7). Unlike *táta* in (3), *babičkem* is a nonword. However, it can be interpreted as consisting of the feminine stem *babičk-* and the masculine singular instrumental suffix *-em*, compatible with the preposition but incompatible with the gender of the stem.[8]

(7) Byl jsem doma s **babičkem** → babičkou
was AUX at home with granny(F).M.SG.INS granny(F).F.SG.INS
'I was at home with Grannie.'

The pattern is shown in Table 3. A more abstract pattern could include only the location and morphology rows.

7 In a fully specified pattern, morphological analysis concerns all relevant categories, including lemma.

8 The bare suffix is ambiguous. It can also express present tense first person plural of some verbal paradigms (*nesem* '[we] carry'). Rather than suggesting such unlikely alternatives, the author is given the benefit of the doubt. For the same reason, we refrain from hypothesizing 'grandpa' (*s dědečkem*) rather than 'granny' (*s babičkou*).

Table 3: The pattern for *babičkem* in (7).

		error annotation	linguistic annotation	
			original	target
location		inflectional suffix	–	–
register		standard	–	–
domain	spelling	two characters' replacement	*em*	*ou*
	morphology	stem/suffix mismatch	stem feminine, suffix masculine	

Tags referring to such patterns can be used as a powerful indicator of the type of interlanguage and the language learner's competence, and can help to build models of interlanguage by machine learning methods. The scheme will be evaluated in trial annotation, including inter-annotator agreement, and tested in machine learning experiments.

Manual annotation can be supported or even replaced by automatic identification of some error types (Jelínek et al. 2012), coupled with a tool suggesting corrections (Ramasamy et al. 2015). Some annotation of a learner corpus can thus be done automatically, without the involvement of human annotators in the process (Rosen 2017).

4 Identifying non-standard language in a corpus

Annotation of word forms and structures in a corpus rarely distinguishes standard language from other varieties. Except for individual word forms in mainstream corpora and error annotation in learner corpora, systematic accounts of non-standard usage are virtually missing. In addition to colloquial, dialectal, obsolete and bookish expressions or imports, described in available lexical resources, non-standard language may also involve performance errors, creative coinages, or emerging phenomena. Most of these phenomena are not covered by standard grammars, but they are still not random, even though the underlying patterns are not easy to discover. In this section, we show an attempt to detect and annotate these phenomena in a treebank/parsebank of Czech.

The theoretical assumption is that linguistic annotation of a corpus represents the meeting point of the empirical evidence (parole) and the theory (langue), in the sense of Saussurean *sign* (de Saussure 1916). Moreover, the annotation is also where multiple levels of analysis and linguistic theories may meet and be explicit about any, even irregular, phenomena. An annotation scheme defined as a formal grammar can help to identify the difference between the regular and irregular, between the language as a system and the use of language.

This is the motivation behind the project of a corpus annotated by standard stochastic tools[9] and checked by a rule-based grammar and valency lexicon, which are also used to infer additional linguistic information about the annotated data.[10] The grammar has the role of a watchdog: to check stochastic parses for both formal and linguistic correctness and consistency. Compliant parses receive additional information: lexical categories receive valency frames to be saturated by complements and project relevant properties to phrasal nodes. Ideally, the grammar should define standard language in the sense of Hirschmann et al. (2007, see section 2 above), although in real life the grammar both overgenerates, leaving some non-standard utterances undetected, and undergenerates, deciding that some standard utterances are not correct.

The grammar consists of a lexical module, providing valency frames, and a syntactic module, checking the parse and projecting information in lexical heads to phrases and complements (dependents). The lexical module, operating on lexical entries derived from external valency lexica, generates available diatheses. The syntactic module matches the generated lexical entries with the data. Categorial information about words and phrases in the data and the lexicon is structured according to a cross-classifying taxonomy, capturing all distinctions present in the standard Czech tagset used in the stochastic parse.[11]

The grammar is implemented in Trale,[12] a formalism designed for grammars based on HPSG, a linguistic theory modeling linguistic expressions as typed feature structures.[13] The grammar differs from a standard implemented HPSG grammar mainly in its role of a constraint solver, rather than a parser or generator. The constraints come from three sources: data, lexicon, and grammar proper. No syntactic rules of the context-free type are needed because the grammar operates on structures already built by a stochastic parser – the syntactic backbone is present in the data, where each sentence has a single parse. Ambiguities or underspecifications may arise only due to the more detailed taxonomy in the treebank format and/or an uncertainty about the choice of a valency frame.

9 See Jelínek (2016).

10 For more detail about the project see, e.g., Petkevič et al. (2015a).

11 See also Petkevič et al. (2015b) for a description of the annotation of periphrastic verb forms using an additional analytical dimension. Periphrastic verb forms are treated with respect to their dual status, i.e. from the paradigmatic perspective as forms of the content verb, and from the syntagmatic perspective as constructions.

12 http://www.ale.cs.toronto.edu/docs/

13 See, e.g., Pollard and Sag (1994) or Levine and Meurers (2006).

The lexical module uses two external valency lexicons: VALLEX[14] and PDT-VALLEX,[15] with their deep valency frames and information about the forms of the syntactic arguments (case, verbal form, etc.). The frames reflect the Praguian valency theory of the Functional Generative Description (Panevová 1994). The lexical module provides the mapping of the frames to their instantiations in specific verbal diatheses and morphological forms, using the same formalism as the syntactic component.

If the syntactic module, after checking the parse using the lexical specifications, decides that the parse complies in all respects, the structure is provided with all available information. If, however, some predicates are left without valency frames, completeness and coherence of the argument structure cannot be checked. Yet some phenomena, such as grammatical agreement, can still be checked. A failure can also be caused by a valency frame. If so, the sentence is additionally checked without that frame. A sentence may also fail due to constraints of the syntactic module. Then the last and weakest test is applied, using only the data format definition without constraints.

Any of these checks may fail due to non-standard linguistic phenomenon in the data, an incorrect decision of the parser or the tagger, or an error in the grammar or lexicon. An efficient and powerful diagnostic is an important task for the future. One option is to make use of the constraint-based architecture by successively relaxing constraints to find the grammatical or lexical constraint and the part of the input responsible for the failure. Another possibility is to use constraints targeting specific non-standard structures or lexical specifications.[16]

Non-standard phenomena can be detected precisely because a grammar of linguistic competence can never fit the corpus as the evidence of linguistic performance completely. To distinguish the cases of truly non-standard language from problems of the grammar on the one hand and to identify and diagnose the types of non-standard language on the other, the diagnostics should be extended to find which specific constraints are violated by which specific words or constructions in the data.

The examples below illustrate the role of the grammar. In (8) and (9) the possessive form agrees in gender and case (and number) with the head noun.

14 See http://ufal.mff.cuni.cz/vallex, Lopatková et al. (2008), Žabokrtský and Lopatková (2007).

15 See Hajič et al. (2003).

16 The so-called *mal-rules* have been used in the context of CALL (computer-assisted language learning) at least by Schneider and McCoy (1998, for users of American Sign Language learning English as their L2), Bender et al. (2004), and Flickinger and Yu (2013) – both implemented in HPSG.

Examples (10) and (11) are different: in (10) the possessive form does not agree with the head noun either in case or in gender, in (11) both in case and gender. Note that the possessive form in (10), which is the same as in (8), does not strike many speakers as incorrect. In the SYN2015 corpus, the share of these non-standard forms is about 4% in the total number of masculine dative singular NPs preceded by the preposition *k*. Example (11) has a similar status, but it is acceptable only to speakers of a dialect of Czech.

(8) Přitiskl se k otcově noze
clung REFL to father's.F.DAT leg(F).DAT
'He pressed against his father's leg.'

(9) Přistoupil k otcovu stolu
approached to father's.M.DAT table(M).DAT
'He appoached his father's table.'

(10) Přistoupil k ?**otcově** stolu
approached to father's.M.LOC/F.DAT table(M).DAT
'He appoached his father's table.'

(11) Přistoupil k ?**otcovo** stolu
approached to father's.N.NOM/ACC table(M).DAT
'He appoached his father's table.'

While (10) and (11) could be seen as examples of suboptimal morphology, (12)–(15) show suboptimal syntax. In (12), an example of *zeugma*, the two coordinated verbs are supposed to share a single object. However, the form of the object (a prepositional phrase) is consistent only with the second verb. In (13), the position of the indirect object of the matrix clause is filled twice: by the headless relative clause and by the personal pronoun. In a standard structure, only a headed relative clause is compatible with an indirect object in the dative case (14). Finally, the matrix clause in (15) includes a subject of the embedded clause (*Gazda*).

(12) ?? Včera jsem viděl a mluvil **s tím člověkem**
yesterday AUX saw and talked with that man
'Yesterday I saw and talked to that man.'

(13) ? **Kdo přijde pozdě**, nic mu nedají
who.nom comes late nothing.ACC him.DAT NEG. give.3.PL
'Who comes late won't get anything.' (intended)

(14)	Tomu,	kdo	přijde	pozdě,	nic	nedají
	that.DAT	who.NOM	comes	late	nothing.ACC	NEG.give.3.PL
	'Who comes late won't get anything.'					

(15) ??	Nebo	já	**Gazda**	nevím,	jak	diktuje
	or	I	Gazda	NEG.know.1.SG	how	dictates
	'Or I don't know how Gazda dictates.'					

In most of the above examples, the stochastic parser ignores the agreement mismatch or the structural anomaly and builds a correct tree. On the other hand, the grammar does not accept the parse, which is the required result. Like every rule-based grammar, it has limited coverage, but a missing account of a phenomenon only means that the grammar overgenerates (is too permissive). Filling gaps in the coverage is another priority for the future.

The grammar and lexicon have been developed and tested on a set of 876 sentences, extracted from the annotation manual of the Prague Dependency Treebank (Hajič et al. 1997), representing a wide range of linguistic phenomena. For 592 sentences a valency frame from the lexicon was found. The number of sentences verified by the grammar is 560. This includes 301 sentences with a valency frame. For more extensive testing, the SYN2015 corpus was used, including about 100 million words, i.e. 7.2 million sentences. For 77% of sentences, at least one valency frame was found and 55% of sentences passed the grammar, 16% including a valency frame, 23% without any valency frame, and 16% after the valency frame was dropped. The next step is to categorize the failures and build a corpus showing the results, including the grammar flags, in a user-friendly way.

5 Conclusion

We have presented two ways to approach non-standard language, with a stress on its proper detection and diagnosis. In the design of an annotation scheme for Czech of non-native learners, we have shown an approach to the analysis of non-standard word forms and structures, based on a layered description of the original and the target expression, combined with corresponding error annotation. In the second study, a method was presented for the detection and diagnosis of non-standard forms and expressions in the grammar-checked annotation of a parsebank. We see this effort as an attempt to tackle a domain of growing importance, one in which the methods and tools available for standard language have only limited usability. Admittedly, we have merely scratched the surface of the topic.

Acknowledgements

This work has been supported by the Grant Agency of the Czech Republic, grant no. 16-10185S.

References

Aharodnik, Katsiaryna, Marco Chang, Anna Feldman and Jirka Hana. 2013. Automatic identification of learners' language background based on their writing in Czech. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJNCLP 2013), Nagoya, Japan, October 2013*, 1428–1436.

Bender, Emily M., Dan Flickinger, Stephan Oepen, Annemarie Walsh and Tim Baldwin. 2004. Arboretum: Using a precision grammar for grammar checking in CALL. In *Proceedings of the InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, Italy.

Berzak, Yevgeni, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza and Boris Katz. 2016. Universal dependencies for learner English. *CoRR*, abs/1605.04278.

Bezuidenhout, Anne L. 2006. Nonstandard language use. In Keith Brown (ed.) *Encyclopedia of Language and Linguistics. Second Edition*, 686–689. Oxford: Elsevier.

Cahill, Aoife. 2015. Parsing learner text: to shoehorn or not to shoehorn. In *Proceedings of The 9th Linguistic Annotation Workshop*, 144–147, Denver, Colorado, USA. Association for Computational Linguistics.

Chomsky, Noam. 1970. Remarks on nominalization. In R. Jacobs and P. Rosenbaum (eds.), *Reading in English Transformational Grammar*, 184–221. Waltham: Ginn and Co.

Dickinson, Markus and Marwa Ragheb. 2015. On grammaticality in the syntactic annotation of learner language. In *Proceedings of The 9th Linguistic Annotation Workshop*, 158–167. Denver, CO.

Dotlačil, Jakub. 2016. Shoda podmětu s přísudkem, pravopis a iluze gramatičnosti. A talk presented at the conference Linguistics and Literary Studies: Paths and Perspectives, Liberec, 22–23 September 2016.

Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera and Holger Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2):139–154. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair.

Eisenstein, Jacob. 2013. What to do about bad language on the internet. In *Proceedings of the North American Chapter of the Association for Computational*

Linguistics (NAACL), 359–369, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Flickinger, Dan and Jiye Yu. 2013. Toward more precision in correction of grammatical errors. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, 68–73, Sofia, Bulgaria. Association for Computational Linguistics.

Hajič, Jan, Jarmila Panevová, Eva Buráňová, Zdenka Urešová and Alla Bémová. 1997. A manual for analytic layer tagging of the Prague Dependency Treebank. Technical Report TR-1997-03, ÚFAL MFF UK, Prague, Czech Republic.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová and Petr Pajas. 2003. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, 57–68. Växjö University Press.

Hirschmann, Hagen, Seanna Doolittle and Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistics structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham.

Jelínek, Tomáš. 2016. Combining dependency parsers using error rates. In *Text, Speech and Dialogue – Proceedings of the 19th International Conference TSD 2016*, 82–92. Springer.

Jelínek, Tomáš, Barbora Štindlová, Alexandr Rosen and Jirka Hana. 2012. Combining manual and automatic annotation of a learner corpus. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karle Pala, editors, *Text, Speech and Dialogue: 15th International Conference*, 127–134. Berlin/Heidelberg: Springer.

Levine, Robert D. and Walt Detmar Meurers. 2006. Head-Driven Phrase Structure Grammar: Linguistic approach, formal foundations, and computational realization. In Keith Brown (ed), *Encyclopedia of Language and Linguistics. Second Edition.* Oxford: Elsevier.

Lopatková, Markéta, Zdeněk Žabokrtský and Václava Kettnerová. 2008. *Valenční slovník českých sloves.* Praha: Karolinum.

Meurers, Detmar. 2013. Natural language processing and language learning. In C. A. Chapelle (ed.), *Encyclopedia of Applied Linguistics*, 4193–4205. Blackwell.

Meurers, Detmar and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning, Special Issue on Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and Interpretation.* Language Learning, 67(S1): 66–95.

Panevová, Jarmila. 1994. Valency frames and the meaning of the sentence. In P. A. Luelsdorff (ed.), *The Prague School of structural and functional linguistics. A short introduction*, 223–243. Amsterdam/Philadelphia: John Benjamins.

Petkevič, Vladimír, Alexandr Rosen and Hana Skoumalová. 2015a. The grammarian is opening a treebank account. *Prace Filologiczne*, LXVII: 239–260.

Petkevič, Vladimír, Alexandr Rosen, Hana Skoumalová and Přemysl Vítovec. 2015b. Analytic morphology – merging the paradigmatic and syntagmatic perspective in a treebank. In Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Hristo Tanev and Roman Yangarber (eds.), *The 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, 9–16, Bulgaria: Hissar.

Plank, Barbara. 2016. What to do about non-standard (or non-canonical) language in NLP. In *KONVENS 2016*.

Plank, Barbara, Hector Martinez Alonso and Anders Søgaard. 2015. Non-canonical language is not harder to annotate than canonical language. In *The 9th Linguistic Annotation Workshop (held in conjunction with NAACL 2015)*, 148–151. Association for Computational Linguistics.

Pollard, Carl J. and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Ramasamy, Loganathan, Alexandr Rosen, and Pavel Straňák. 2015. Improvements to Korektor: A case study with native and non-native Czech. In Jakub Yaghob, editor, *ITAT 2015: Information technologies – Applications and Theory / SloNLP 2015*, 73–80, Charles University in Prague.

Rosen, Alexandr. 2014. A 3D taxonomy of word classes at work. In Ludmila Veselovská and Markéta Janebová, editors, *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*, volume 4 of *Olomouc Modern Language Series*, 575–590. Olomouc: Palacký University.

Rosen, Alexandr. 2017. Introducing a corpus of non-native Czech with automatic annotation. In Piotr Pęzik, Jacek Waliński, and Krzysztof Kosecki, editors, *Language, Corpora and Cognition*, 163–180. Frankfurt am Main, Bern, Bruxelles, New York, Oxford, Warszawa, Wien: Peter Lang.

Rosen, Alexandr, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation – Special Issue: Resources for language learning*, 48(1): 65–92.

de Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Paris. Publié par Ch. Bally et A. Sechehay avec la collaboration de A Riedlinger.

Schneider, David and Kathleen F. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Volume 2, ACL '98, 1198–1204, Stroudsburg, PA, USA. Association for Computational Linguistics.

Žabokrtský, Zdeněk and Markéta Lopatková. 2007. Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, (87): 41–60.

Renate Raffelsiefen, Anja Geumann

Phonological Analysis at the Word Level: The Role of Corpora

Abstract Notions such as "corpus-driven" versus "theory-driven" bring into focus the specific role of corpora in linguistic research. As for phonology with its intrinsic focus on abstract categorical representation, there is a question of how a strictly corpus-driven approach can yield insight into relevant structures. Here we argue for a more theory-driven approach to phonology based on the concept of a phonological grammar in terms of interacting constraints. Empirical validation of such grammars comes from the potential convergence of the evidence from various sources including typological data, neutralization patterns, and in particular patterns observed in the creative use of language such as acronym formation, loanword adaptation, poetry, and speech errors. Further empirical validation concerns specific predictions regarding phonetic differences among opposition members, paradigm uniformity effects, and phonetic implementation in given segmental and prosodic contexts. Corpora in the narrowest sense (i.e. "raw" data consisting of spontaneous speech produced in natural settings) are useful for testing these predictions, but even here, special purpose-built corpora are often necessary.

Keywords Speech corpora, German vowels, phonological grammar, abstractness, Optimality Theory

1 Introduction

Phonology is concerned with capturing the contrastive potential of a language, aiming at a comprehensive account of the ways in which differences in meaning can be conveyed through sound differences. Traditionally, a phonological description includes an inventory of phonemes, organized in terms of oppositions or distinctive features, along with rules for the combination and prosodic organization of the phonemes. Such a description then determines the lexical phonemic representations of words, which form the input to phonetic implementation.

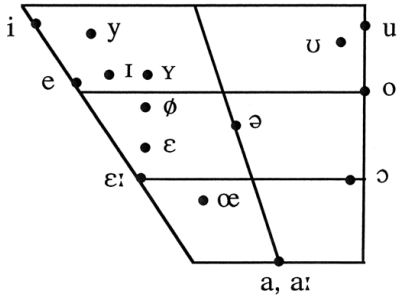
The key intuition guiding phonemic analyses concerns a basic classification of linguistic material in terms of sameness versus distinctness, focusing on conditions for determining whether or not

– phonetically distinct sounds represent the same phoneme

– phoneme pairs represent the same (i.e. "proportional") opposition

The answer to the first question again crucially refers to the notion of sameness since proof of phonemic distinctness presupposes the occurrence of distinct sounds in identical contexts. Applying this condition to German typically results in an inventory of fifteen or more vowel phonemes, which are then investigated and associated with IPA-symbols. Two descriptions with vowels arranged in accordance with IPA-conventions, one proposed by Kohler (1999: 87), see (1a), the other by Eckert & Barry (2005: 111), see (1b), are shown below.

(1) a.



b.

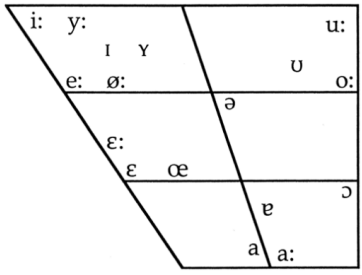


Figure 1: a. The German vowel system according to Kohler (1999). b. The German vowel system according to Eckert and Barry (2005).

While Eckert & Barry posit a vowel /ɐ/ to represent the unstressed syllable in words like *Vater* 'father', Kohler apparently considers that sound the same as other independently established phonemes. There is agreement that two vowel pairs differ in quantity only (i.e. /a/-/aː/ as in *prallen* 'to bump' - *prahlen* 'to boast', /ɛ/-/ɛː/ as in *stellen* 'to put' - *stählen* 'to steel'), in contrast to all other oppositions, which are deemed to involve no phonemic quantity contrast (cf. 1a) or one linked to quality contrasts (cf. 1b). More radically different assessments of vowel sameness are seen in the works of others, including the view that there are no more than eight distinct vowels in German (Vennemann 1991, Becker 1998).

How can corpora help decide among such phonemic analyses or help evaluate the merits of abstract representation in general? Is there hope that ever larger corpora of spontaneous speech, subjected to ever more precise measurements and ever more sophisticated statistical modeling, could further our understanding of phonemic structure? How can quantitative methods capture the notion of

phonemic sameness, which is rooted in the intuition that physical differences are abstracted away from and items are classified the same as long as those differences can be attributed to context? How can such methods capture abstractions in the minds of speakers which clearly are not amenable to direct measurement?

The approach to pinpointing phonemic structure to be illustrated below is rooted in the idea of a phonological grammar as a language-specific ranking of universal constraints (Prince & Smolensky 1993). While phonemic structure and the concept of abstractness are rarely addressed in such frameworks, we will argue that the interaction of constraints and their inherent properties yield insight into such structure. On this approach the focus shifts to data resources that shed light on constraints and their effects on phonological structure. Empirical support comes from the convergence of various types of independent evidence.

The paper is organized as follows. Section 2 presents some basic claims of constraint-based grammars, illustrating these with the role of roundedness in the vowel system of German. Section 3 focuses on the relevance of constraints in distinguishing between phonemic and subphonemic structure, to be illustrated with length versus quality differences in German vowels. Section 4 discusses some of the currently existing resources.

2 Constraint-based grammar: some basic ideas

Optimality Theory envisions phonological grammar as language-specific resolutions of conflicts among universal constraints (Prince & Smolensky 1993). The core conflict concerns the desirability to maximize contrast, by allowing all types of structure to distinguish morphemes, versus the desirability to minimize phonological markedness, to enhance ease of production and perception. Additional constraints concern correspondence of structure among words, requiring sameness of structure both at the syntagmatic level, to enhance cohesion (e.g. rhymes, alliteration), and at the paradigmatic level, to minimize allomorphy and enhance recognition of paradigmatic relatedness.

To illustrate a language-specific resolution of the core conflict between the maximization of potential contrast and satisfaction of markedness constraints, consider the roundedness contrast in German in (2). The stressed vowels are represented without duration marks, as duration will be argued to be a subphonemic property in German (cf. section 3).

(2) a. /'ʃpilən/ <spielen> ‘to play’
/'kɪsən/ <Kissen> ‘pillow’
/'lezən/ <lesen> ‘to read’
/'kɛnən/ <kennen> ‘to know’

b. /'ʃpylən/ <spülen> ‘to rinse’
/'kʏsən/ <küssen> ‘to kiss’
/'løzən/ <lösen> ‘to solve’
/'kœnən/ <können> ‘to be able to’

Contrastiveness as in (2) motivates the assumption of an active faithfulness constraint FAITH([±round]). Formally, such a constraint concerns the relation between an input and the corresponding output, requiring the "faithful" preservation of the input structure. To maximize potential contrast, it would be ideal if roundedness were contrastive for all vowels, including low and back vowels. The restriction of this contrast to the vowel pairs illustrated in (2) indicates a specific interaction among FAITH([±round]) and phonological markedness constraints prohibiting the cooccurrence of the feature [±round] with other features (e.g. *V{[+back][-round]} (Back unrounded vowels are prohibited), *V{[-back][+round]} (Front rounded vowels are prohibited)). The ranking in (3) says that in German for back and low vowels, it is more important to satisfy the relevant markedness constraints than to exploit the contrastive potential of lip roundedness. Only for non-low front vowels is the potential for contrast valued more than the satisfaction of the relevant markedness constraint (i.e. *V{[-back][+round]}). (Constraint domination is marked by the symbol ">>".)

(3) *V{[+back][-round]}, *V{[+low][+round]} >> FAITH(V[±round]) >> *V{[-back][+round]}

Phonological markedness constraints are presumably ultimately grounded in phonetics, expressing relative difficulties in articulating or perceiving certain structures compared to others (e.g. specific coordinations between tongue positions and lip roundedness). They are reflected in asymmetries in the distribution of sounds in the languages of the world documented in databases such as UPSID[1], which is based on 317 languages. Links between lip roundedness and tongue advancement are shown by the fact that 94 % of front vowels are unrounded whereas 93.5 % of back vowels are rounded (Maddieson 1984: 124). Among the vowels classified as low central monophthongs in the languages in question, 392 unrounded compare to a single rounded vowel (Maddieson 1984: 124).

A representation of phonological grammar in terms of rankings among universal constraints as in (3) is superior to a mere listing of phonemes in that it relates the actual to the potential. Such a model predicts that more marked structure (e.g. rounded front vowels in German) implies the existence of the corresponding less marked structure (e.g. unrounded front vowels in German). This is because there is no ranking of independently motivated markedness constraints which would describe a language where marked structures exist to the exclusion of the corresponding less marked structures.

1 This acronym stands for *UCLA Phonological Segment Inventory Database* (Maddieson & Precoda 1990). For more discussion see section 4.

To verify the existence of the respective less marked phonemes it is necessary to establish the relevant relations and to demonstrate the presence of consistent phonetic correlates. The relations in question are supported by correspondence patterns, including regular sound alternations in paradigms[2] and also so-called impure rhymes, which are characterized by specific relaxations of a general requirement for sameness. Consider the German word pairs in (4)[3], which function as rhymes despite the difference in vowel roundedness. These rhymes then support the specific phoneme correspondences illustrated by the minimal pairs in (2).

(4) /y/ : /i/ *grüßen* 'to greet' – *fließen* 'to flow'
/ʏ/ : /ɪ/ *Sünder* 'sinner' – *Kinder* 'children'
/ø/ : /e/ *schön* 'beautiful' – *stehn* 'to stand'
/œ/ : /ɛ/ *Töchter* 'daughters' – *Wächter* 'guard'

Reference to the feature [±round] in the grammar stated in (3) to capture the vowel opposition illustrated in (2) is motivated by the relevance of the respective markedness constraints. A consistent phonetic difference is confirmed by observing the degree of lip roundedness during the articulation of the vowels in each pair in (4). However, not all phonetic reflexes are easily assessed on an introspective basis and in general there are many advantages to conducting phonetic studies based on acoustic measurements. Such studies concern the resonances, known as formants, which change according to the size and the shape of the vocal tract thereby reflecting on articulatory properties (Peterson & Barney 1952). For example, the first formant frequency (F1) increases as the tongue lowers. F1 decreases, while the second formant frequency (F2) increases, as the tongue body advances. All formant frequencies, especially F2 and F3, decrease with increased lip roundedness as a result of the concomitant elongation of the vocal tract (Hixon et al. 2008).

Regarding the pairs in (4), there is accordingly a prediction that for each unrounded vowel, the values for F2 and F3 should be higher than those for the corresponding rounded vowels. This prediction is borne out by the measurements of the relevant vowel formants based on recordings of 26 female speakers in the Kiel Corpus of Read Speech (cf. section 4).[4]

2 Correspondence involving paradigmatic relations can be illustrated with plural-singular pairs (e.g. /ʃtylə/ <Stühle> 'chairs' – /ʃtul/ <Stuhl> 'chair', /flʏsə/ <Flüsse> 'rivers' – /flʊs/ <Fluss> 'river'), which confirm the existence of a less marked rounded back vowel corresponding to each more marked front rounded vowel.

3 These rhymes are adopted from the poem "Romanzen vom Rosenkranze" by Clemens Brentano.

4 Formant values were extracted automatically with PRAAT (Boersma & Weenink (2016)) at 50% of the vowel duration. The numbers of tokens for individual vowels (stressed and unstressed) are as follows: /i/ 1215, /y/ 289, /ɪ/ 2,536, /ʏ/ 264, /e/ 978, /ø/ 149, /ɛ/ 1,070, /œ/ 245.

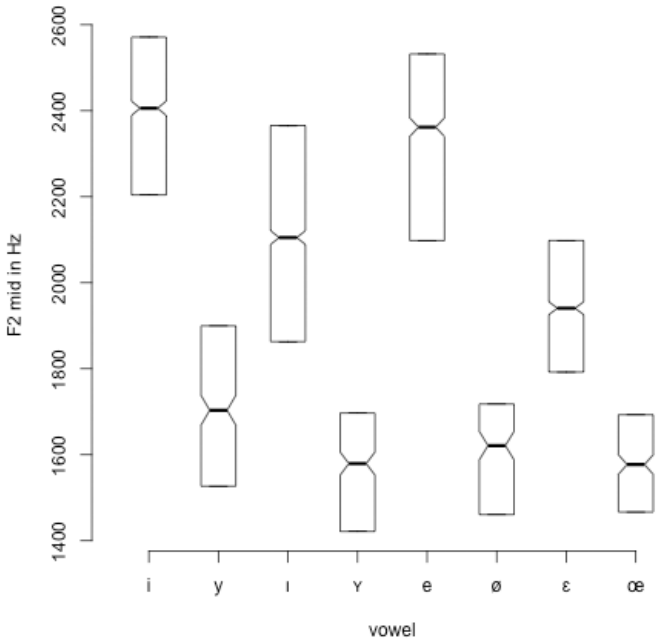


Figure 2a: *Kiel Corpus* 26 f speakers, Formant F2 values in Hz.

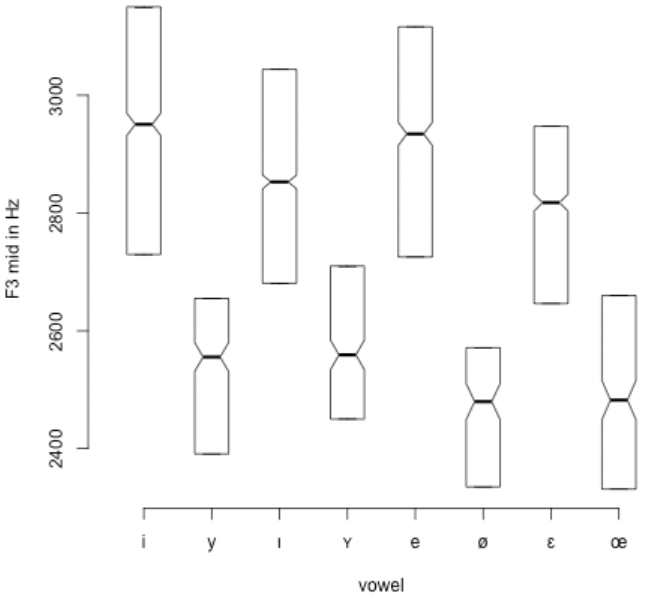


Figure 2b: *Kiel Corpus* 26 f speakers, Formant F3 values in Hz.

The boxplots in Figure 2 show non-overlapping indentations for all relevant pairs (e.g. /i/–/y/), which means that the median values differ significantly (Chambers et al 1983).[5]

A comparison of the four vowel pairs in Figure 2 shows that the respective differences among the formant values differ considerably. For instance, the pair /e/–/ø/ exhibits a larger difference among the values for both F2 and F3 than the pair /ɛ/–/œ/. Such disparities are consistent with the assumption of a single phonological opposition as long as they can be attributed to independent differences (e.g. larger difference among F2 values in pairs of peripheral vowels (/e/–/ø/, /i/–/y/) compared to the corresponding pairs of centralized vowels (/ɛ/–/œ/, /ɪ/–/ʏ/)). The analysis of all of the relevant pairs as instances of a single phonological roundedness opposition is expressed in terms of positing a single faithfulness constraint FAITH(V[±round]) and its interaction with other constraints as in (3).

The claim that the constraint ranking in (3) captures the role of roundedness in German phonology is supported by independent evidence concerning historical change.

Here again, we find an asymmetry to the effect that an increase of markedness (the emergence of rounded front vowels) comes about through context-sensitive change whereas context-free change consistently leads to a decrease of markedness (unrounding of front vowels). This generalization can be illustrated with the development of the English verb *kiss* in (5), where an increase in markedness (/ʊ/ => /ʏ/) results from assimilation (fronting of /ʊ/ to agree with the following front vowel /i/). The subsequent loss of rounding in front vowels (/ʏ/ => /ɪ/) is context-free and reduces segmental markedness:

(5) Old Saxon *kussian* > Old English *cyssan* > Modern English *k*/ɪ/*ss* <kiss>

Additional sources of front rounded vowels in German are illustrated in (6).[6] The sporadic changes from less marked to more marked vowels invariably involve segmental contexts consisting of labial fricatives [v], [f] or [ʃ][7], all of which favor the perception of a rounded vowel:

(6) MHG *wirde* > NHG *W*/ʏ/*rde* <Würde> 'dignity'
MHG *vinf* > NHG *f*/ʏ/*nf* <fünf> 'five'
MHG *zwelf* > NHG *zw*/œ/*lf* <zwölf> 'twelve'
MHG *lewe* > NHG *L*/ø/*we* <Löwe> 'lion'
MHG *leschen* > NHG *l*/œ/*schen* <löschen> 'to extinguish'

5 Outliers are not presented in the boxplots but are included in the calculations.

6 The changes are sporadic as unrounded vowels are often preserved in the contexts in question (e.g. NHG *W*/ɪ/*rbel* <Wirbel> 'whirl', NHG *W*/ɛ/*lle* <Welle> 'wave').

7 [ʃ] is pronounced with strongly protruded lips in German (cf. Wängler 1964).

In other contexts, changes involving roundedness consistently favor unmarked unrounded front vowels. The following changes concern vowels spelled with the grapheme <y>, which is historically linked to rounded /y/ or /ʏ/, but, unlike the grapheme <ü>, also associates with unrounded vowels in German (s. Dudenband 6: 913).

(7) G/ʏ/mnásium > G/ɪ/mnásium <Gymnasium> 'secondary school'
s/ʏ/mpátisch > s[ɪ]mpátisch <sympathisch> 'likable'
S/ʏ/stém > S[ɪ]stém <System> 'system'

The asymmetry in historical change illustrated above is predicted by the grammar in (3) if one were to assume inputs consisting of actual word forms encountered by hearers. Faithfulness constraints would then make their force felt only if a given sound property has been perceived. Otherwise markedness prevails and the unmarked segments will emerge. This approach also makes sense of the fact that reanalysis to unmarked vowels as in (7) is more common in unstressed positions because stressed syllables favor the perception of contrasts (cf. the stability of roundedness in words like '*P*/y/*thon* <Python> 'python', '*G*/y/*ros* <Gyros> 'gyros'). The connection in question can be expressed by way of linking faithfulness constraints to prominent positions (e.g. $\text{FAITH}_{\text{stress}}$) and by imposing a universally fixed ranking to the effect that $\text{FAITH}_{\text{POS}}$ (POS = "prominent position") dominates the corresponding general faithfulness constraint. This phenomenon, known as "positional faithfulness" (Beckman 1998), is also relevant to the analysis of speech errors illustrated in (8)[8], which appear to favor the alignment of marked structures with prominent positions. The correct and presumably intended forms are given in parenthesis.

(8) M[ɪ]s't[ø:]rium (M[ʏ]s't[e:]rium <Mysterium> 'mystery')
S[ɪ]n't[ø:]se (S[ʏ]n't[e:]se <Synthese> 'synthesis')
S[ɪ]l'v[œ]ster (S[ʏ]l'v[ɛ]ster <Sylvester> 'New Year's Eve')
Di[e]'z[ø:]se (Di[ø]'z[e:]se <Diözese> 'diocese')
Z[i]'l[ʏ]nder (Z[y]'l[ɪ]nder <Zylinder> 'cylinder')

A phonological grammar in terms of ranked constraints as in (3) accounts for both the distribution of phonemes, thus capturing potential contrast, and the stability of phonological structure. Significantly, such a grammar provides clear guidance for research based on annotated speech corpora, singling out specific

8 The examples in (8) stem from personal communication, published speech error collections (Leuninger 1996), or common misspellings in internet data (e.g. *Zilynder*, *Silvöster*).

sound structures for comparison and focusing the investigation on the question of how certain abstract structures are implemented in various segmental and prosodic contexts. For instance, the juxtaposition of the measurements shown in Figure 2 indicates closer F2 and F3 values for roundedness contrasts involving centralized vowels, which may account for the higher rate of phonemic reanalysis for such vowels (cf. MND *flistern* > '*fl*[ʏ]*stern* <flüstern> 'to whisper', but MND *vlise* > '*Fl*[i:]*se* <Fliese> 'tile').

The data reviewed so far illustrate types of evidence to support grammatical descriptions in terms of interacting constraints as well as the use of speech corpora to verify the presence of consistent phonetic correlates. The following section illustrates ways in which evidence from constraint interactions can resolve questions concerning phonemic abstractness along with additional ways in which acoustic studies could verify such analyses.

3 Identifying phonemic oppositions

While there is a consensus that the minimal pairs listed in (2) illustrate a single rounding opposition, other cases raise substantial controversy. Recall the lack of consensus regarding the role of quantity versus quality in the analysis of German vowels addressed above. A complete list of relevant opposition members, represented phonetically in square brackets and referred to as "A-vowels" versus "B-vowels" for now, is illustrated in (9). The cases which have been claimed to involve a pure quantity opposition are listed in (9b), where the symbols /a:/ and /ɛ:/ presented in the charts in (1) are replaced by symbols indicating quality differences (i.e. [ɑ:] and [e*:])

		A-vowels	B-vowels
(9)	a.	/m[i:]nə/ <Mine> 'mine'	/m[ɪ]nə/ <Minne> 'love'
		/d[y:]nə/ <Düne> 'dune'	/d[ʏ]nə/ <Dünne> 'thinness'
		/b[u:]lə/ <Buhle> 'paramour'	/b[ʊ]lə/ <Bulle> 'bull'
		/d[o:]lə/ <Dohle> 'jackdaw'	/d[ɔ]lə/ <Dolle> 'rowlock'
		/h[ø:]lə/ <Höhle> 'cave'	/h[œ]lə/ <Hölle> 'hell'
		/ʃt[e:]lən/ <stehlen> 'to steal'	/ʃt[ɛ]lən/ <stellen> 'to put'
	b.	/ʃt[e*:]lən/ <stählen> 'to steel'	?/ʃt[ɛ]lən/ <stellen> 'to put'
		/pʁ[ɑ:]lən/ <prahlen> 'to boast'	/pʁ[a]lən/ <prallen> 'to bump'

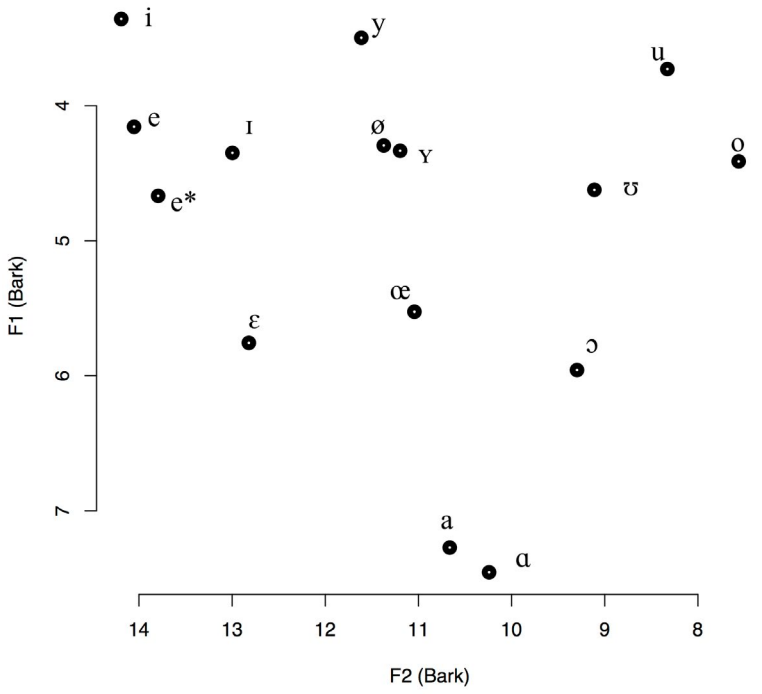


Figure 3a: Vowel chart F1/F2 plane in Bark for stressed German vowels. *Kiel Corpus of Read Speech*, 26 female speakers.

Measurements of F1 and F2 for these vowels, again based on the female speakers of the Kiel Corpus, are given in Figure 3a. The respective values for duration are listed in Figure 3b.[9]

The values in Figure 3 are largely consistent with both phonemic analyses indicated in (1). It is doubtful that additional measurements, based on larger corpora, could answer the question of whether the length contrasts are phonemic for all, some, or no pairs. Indeed, none of the phonetic studies considered so far seem to offer a clear basis for deciding which vowels form opposition members in the first place. Proximity of positions within the formant charts alone is hardly decisive as for instance the vowels in /d[ʏ]nə/ <Dünne> 'thinness' versus /h[ø:]lə/ <Höhle> 'cave' are represented with distinct symbols in all descriptions known to us, despite exhibiting greater similarity than any of those in (9b).

9 For our calculations we used the Burg algorithm, searching for 5 formants in the range from 0-5500 Hz for females. The number of tokens, all of them stressed, are as follows: /a/ 1,157, /ɑ/ 575, /ɛ/ 698, /e/ 619, /e*/ 36, /ɪ/ 645, /i/ 419, /ɔ/ 279, /o/ 231, /œ/ 81, /ø/ 138, /ʊ/ 324, /u/ 365, /ʏ/ 209, /y/ 205.

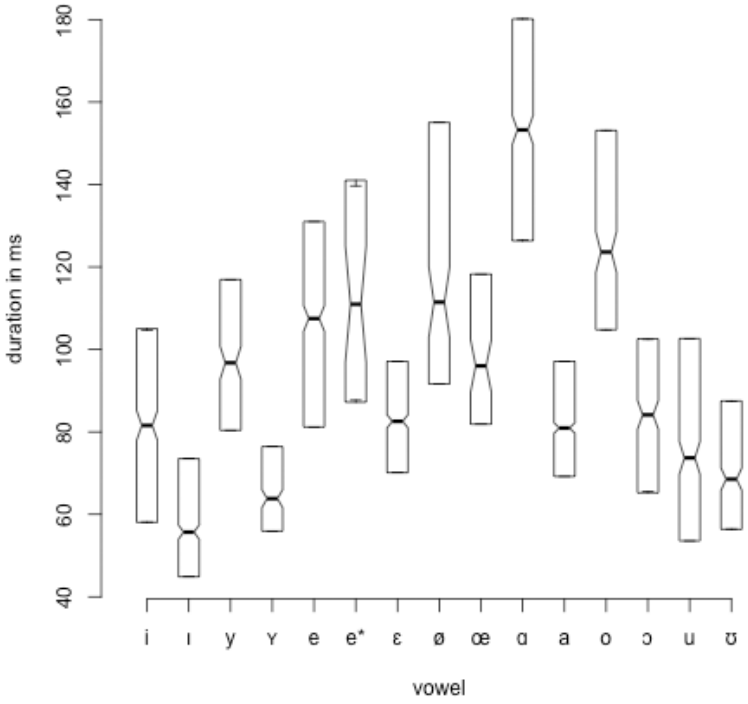


Figure 3b: Vowel durations in ms for stressed German vowels.
Kiel Corpus of Read Speech, 26 female speakers.

Below we will briefly indicate how a constraint-based approach may resolve these questions, focusing on the sort of data resources needed for establishing constraint interactions. The question of whether or not the pairs in (9) form a single opposition is addressed in section 3.1, while arguments for separating phonemic from subphonemic structure to identify that opposition are addressed in section 3.2. Arguments for identifying respective opposition members are reviewed in 3.3. Section 3.4 discusses corpus-based acoustic studies relevant to verifying the results.

3.1 Establishment of a single opposition

The analyzability of all vowel pairs in (9) as a single opposition depends on whether there are parallel restrictions indicative of single constraint interactions. The investigation focuses then on neutralization patterns, to establish the existence of contexts where all A-vowels can appear, to the exclusion of all B-vowels,

and vice versa. One such context is given in (10a), as all A-vowels, but no B-vowels, occur before another syllabic vowel.

(10) a.	/'n[ɑ:]ə/ <nahe> 'near'	b.	<< OHG nāh
	/'[e:]ə/ <Ehe> 'marriage'		<< OHG ēwa
	/'ʁ[u:]ə/ <Ruhe> 'quiet'		<< OHG ruowa
	/'m[y:]ə/ <Mühe> 'effort'		<< OHG muohi
	/'dʁ[o:]ən/ <drohen> 'to threaten'		<< OHG drouwen
	/'ʁ[i:]o/ <Rio> place name		Spanish [rrío]

The demonstration of systematic restrictions on phonological form is inherently problematic as it may seem to require an exhaustive examination of all relevant data. In addition, there is a possibility that the absence of specific patterns is synchronically accidental, caused by the imitation of the given and ultimately resulting from historical circumstances. Such conditions might fully account for the restrictions on the prevocalic vowels observed in (10a) as they go back to long vowels or diphthongs in Old High German (OHG) shown in (10b). Also in loan words the relevant structure could exist independently in the source language, adapted "faithfully" by the borrowers, without necessarily being represented in their phonological grammar.

There is a question then of which types of data are best suited to reveal genuine phonological restrictions caused by active phonological markedness constraints. All data involving potential modification of observable input structures are ideal as such modifications necessarily indicate the dominance of markedness constraints over faithfulness. Apart from cases of historical change and speech errors discussed above, the most significant sources include acronyms and the adaptation of loan words. The latter type is illustrated in (11), where apparent B-vowels in prevocalic position in the French source words are systematically replaced by A-vowels in German.[10]

(11) French /kl[ɔ]'ak/ <cloaque> 'sewer' => German /kl[o]'ɑkə/ <Kloake> 'sewer'
French /n[ɔ]'ɛl/ <noël> 'Christmas' => German /n[o]'ɛl/ <Noël> 'French Christmas carol'
French /p[ɔ]e'zi/ <poésie> 'poetry' => German /p[o]e'zi / <Poesie> 'poetry'

10 The data in (11) raise a question concerning the status of the respective input and output forms. French acoustic forms could be mapped to forms perceived by German learners. Alternatively, French structures perceived by German speakers could be mapped to outputs they produce in speech. Either view involves modifications which presuppose an active markedness constraint.

Reference to loan words as a source for revealing active markedness constraints is potentially complicated by instances of so-called unassimilated loans illustrated in (12), which are characterized by special efforts on the parts of speakers to closely imitate the pronunciations of words in the respective donor languages.

(12) [nɔ'e] <Noé> French male given name
[nɔ'ɑ:j] <Noailles> French place name

Such cases are best set aside as long as they are limited to a few uncommon words (e.g. proper names) and/or are characterized by varying pronunciations. Conceivably, individual words can be imitated as wholes, with no impact on the phonological grammar. In general, the lack of modification of structure seen in unassimilated loan words does not contradict the assumption of active markedness constraints, but cases of systematic modification strongly support that assumption.

The second type of data mentioned above, acronyms, is also characterized by systematic restrictions on output forms which cannot be due to imitation of given forms. The data in (13) illustrate again the systematic exclusion of B-vowels in prevocalic position.

(13)	/'tˢ[o:]atˢ/	ZOAZ	Zentrales ([ɔ])/Organisations- und Abrechnungszentrum
	/'f[e:]ap/	VEAB	Volkseigener ([ɛ])/Erfassungs- und Aufkaufbetrieb für landwirtschaftliche Erzeugnisse
	/'f[i:]ak/	VIAG	Vereinigte ([ɪ])/Industrie-Unternehmen AG
	/'ʀ[i:]as/	RIAS	Rundfunk ([ɪ])/im amerikanischen Sektor

The restriction to the A-vowels marked in the acronyms in (13) can be linked neither to the vowels contained in the relevant source words (cf. the right-hand column in (13)), nor to conventions concerning grapheme-phoneme correspondence[11]. Although acronym formation or the adaptation of loan words may appear to be marginal phenomena, both are associated with highly regular modifications of sound structure which can be explained only by active phonological markedness constraints. A consistent convergence seen in the relevant output restrictions (e.g. the restriction to A-vowels in prevocalic position in both /kl[o]'ɑkə/ <Kloake> 'sewer' and /'tˢ[o:]atˢ/ <ZOAZ>) documented for all relevant opposition pairs may indeed suffice to establish systematic gaps in the distribution of phonemes, obliterating the need for an exhaustive examination of dictionaries.

11 For instance, the grapheme <E> associates with a lax vowel in /pɛk/ <PEG> (based on *Perkutane endoskopische Gastrostomie*) versus a tense vowel in /pekɪp/ <PEKIP> (based on *Prager Eltern-Kind Programm*).

As for the overall distribution of A- versus B-vowels in German, a thorough investigation of neutralization patterns indicates strictly parallel patterns within each class. For instance, the restriction to A-vowels, which are "tense" and phonetically long, in prevocalic position in (13) also extends to low vowels illustrated in (14):

(14) /'l[ɑ:]ɔs/ Laos (possibly adopted from French [la'o:s] <Laos> 'Laos'
/'tˢ[ɑ:]ɛt/ ZAED Zentralstelle für Atomenergie-Dokumentation

The exclusion of all phonetically short B-vowels in the stressed prevocalic position supports the presence of a single opposition. Many additional contexts can be found, where either only A-vowels occur, to the exclusion of all B-vowels or only B-vowels occur, to the exclusion of all A-vowels.[12] This parallelism strongly argues in favor of a single opposition distinguishing A- versus B-vowels, not a mixed system as suggested by Kohler's depiction in (1a).

3.2 Identifying the nature of the opposition

As was noted above, the assumption of phonological markedness constraints rests on cross-linguistic asymmetries in the distribution of sounds. Their proper identification in individual languages is accordingly determined primarily by the overall neutralization patterns. As for the opposition of A- versus B-vowels in German, the observed restrictions suggest reference to syllable structure, invoking markedness constraints of the type "No B-vowels in open syllables", "No A-vowels in closed syllables". This particular context is consistent with a quality contrast, as is shown by the so-called *Loi de Position* in French, which bans vowels in word-final open versus closed syllables based strictly on their quality, regardless of length (e.g. "/o/ and /ø/, but no /ɔ/ or /œ/, in open syllables", "/ɛ/, but no /e/ in closed syllables"). However, in general the syllable structure contexts in question may also be consistent with a quantity opposition, provided that rules known as "Open Syllable Lengthening" and "Closed Syllable Shortening" can in fact be shown to be neutralizing.

As for German, the syllable-based restrictions in question appear to target quality rather than quantity. This is because the relevant neutralization patterns are also observed in unstressed position, where all vowels are short (cf. the data in (11)). Moreover, there are additional neutralization patterns clearly betraying

12 The restriction to only B-vowels is for instance seen before sonorant-obstruent clusters which include a non-coronal segment (e.g. /'v[ɔ]lkə/ (*/'v[o:]lkə/) <Wolke> 'cloud', /'f[a]lkə/ (*/'f[ɑ:]lkə/) <Falke> 'falcon').

reference to quality rather than quantity. The words in (15a) illustrate a restriction to B-vowels before a closed syllable containing a corresponding B-vowel, an apparent harmony effect as vowels before similar consonants in (15b) are not affected (cf. Raffelsiefen 2016).

(15) a. /'b[ɔ]tʀɔp/ <Bottrop> 'place name'
/'n[ʊ]bʊk/ <Nubuk> 'nubuck'
/'v[ɪ]tɪp/ <witib> 'widow'
b. /'l[oː]tʀɪŋən/ <Lothringen> 'Lorraine'
/'t[uː]bɑ/ <Tuba> 'tuba'
/'v[iː]tɑ/ <Vita> 'vita'

Evidence for an active harmony constraint in (15a) is highly significant as harmony is known to universally refer only to quality features, never to length. The data in (16a) illustrate the relevant harmony effect for low vowels, further demonstrating the parallel behavior of all A- versus B-vowels. The acronyms in (16b) show the synchronic productivity of the relevant restrictions.[13]

(16) a. /'t[a]bak/ <Tabak> 'tobacco' /'ʀ[ɑː]bə/ <Rabe> 'raven'
/'m[a]dras/ <Madras> 'place name' /'p[ɑː]dʀə/ <Padre> 'padre'
b. /'h[a]pak/ <HAPAG> /'[ɑː]po/ <APO>
/'t[a]kraf/ <TAKRAF> /'n[ɑː]gʀɑ/ <NAGRA>

Given the necessary reference to quality features to capture the neutralization patterns in (15) and (16), as opposed to the absence of cases where reference to quantity is needed to capture potential contrast patterns, the opposition referred to as A- versus B-vowels in (9) can be analyzed as a fundamental quality opposition. Vowel length constitutes then a subphonemic property. Additional research, including investigations of cross-linguistic patterns, is needed to properly identify the quality feature in question. Here, we tentatively choose the feature [± peripheral] (Lindau 1978). The restriction of the relevant contrast to stressed syllables indicates an active positional faithfulness constraint $\text{FAITH}(\pm\text{PER})_{\text{STRESS}}$ whose interaction with various markedness constraints captures the distribution of peripheral versus centralized vowels in German (Raffelsiefen 2016). Section 3.4 focuses on studies based on speech corpora suited to the empirical testing of the analysis.

13 HAPAG: Hamburg-Amerikanische Paketfahrt-Aktien-Gesellschaft, TAKRAF: Tagebauausrüstungen, Krane und Förderanlagen, APO: Außerparlamentarische Opposition, NAGRA: Nationale Genossenschaft für die Lagerung radioaktiver Abfälle.

3.3 Identifying corresponding opposition members

As was noted above, the identification of individual opposition members is supported by evidence pertaining to violations of strict correspondence constraints pertaining to both paradigmatic and syntagmatic relations. Paradigmatic correspondences are illustrated in (17), where a centralized vowel in an unstressed closed syllable alternates with a peripheral vowel in an unstressed open syllable. The vowel alternation is caused by a vowel-initial suffix carrying main stress, which conditions the syllabification of the preceding consonant as an onset, as opposed to the coda syllabification of the corresponding consonant in the base. Stresslessness is crucial as stressed vowels exhibit regular paradigm uniformity effects (see 3.4). The derived formations in (17b) are marked with question marks because they are not attested.[14]

(17) a. *Ják*[ɔ]*b* <Jakob> 'male name' *Jak*[o.]*bíner* <Jakobiner> 'Jacobin'
Tíb[ɛ]*t* <Tibet> 'Tibet' *Tib*[e.]*táner* <Tibetaner> 'Tibetan'
Lím[ɪ]*t* <Limit> 'limit' *lim*[i.]*tíeren* <limitieren> 'to limit'
Sább[a]*t* <Sabbat> 'Sabbat' *Sabb*[ɑ.]*tíst* <Sabbatist> 'sabbatist'
b. *Kál*[ʏ]*m* <Kalym> 'kalym' ?*kal*[y.]*míeren*
Báf[œ]*g* <Bafög> 'funding for students' ?*baf*[ø.]*gíeren*

The evidence from the paradigmatic alternations in (17) agrees with evidence pertaining to rhyme. The examples for assonance in (18) exhibit identical values for all contrastive vowel features other than [±peripheral].[15]

(18) '[ʊ]*nter* <unter> 'under' – '*Gr*[u:]*be* <Grube> 'pit'
'*S*[ʏ]*nde* <Sünde> 'sin' – '*w*[y:]*hlen* <wühlen> 'to rummage'
'*f*[a]*ngen* <fangen> 'to catch' – '*gr*[ɑ:]*ben* <graben> 'to dig'
'*tr*[ɛ]*ffen* <treffen> 'to meet' – '*L*[e:]*hrer* <Lehrer> 'teacher'
'*M*[ɛ]*sser* <Messer> 'knife' – '*Tr*[e*:]*nen* <Tränen> 'tears'

Assuming that the stressed vowels in the examples *Tränen* and *Lehrer* cited in (18) are indeed distinct, the assonance patterns support the correspondence relations indicated in (9), where both of the peripheral vowels in question correspond to centralized /ɛ/.

14 The relevant alternations ought to also be tested experimentally, ideally with illiterate speakers to exclude possible correspondence effects pertaining to graphemes.

15 The examples in (18) are also adopted from Brentano's poem "Romanzen vom Rosenkranze".

3.4 Verifying phonological analyses

The establishment of a single quality opposition for the vowel pairs illustrated in (9) predicts the presence of a consistent phonetic correlate. The measurements in Figure 4 are based on all 15 vowels in stressed position pronounced by female speakers in the *Kiel Corpus*[16] and demonstrate that each A-vowel is more peripheral than the corresponding B-vowel. In particular, it is shown that for a specific central position the peripheral vowel is always further away than the corresponding centralized vowel. The central position is calculated individually as the mean F1 and F2 value for all relevant vowels in a given (sub)corpus. The distance is then calculated as the Euclidian distance to the central position for each individual vowel in the F1 by F2 vowel space.

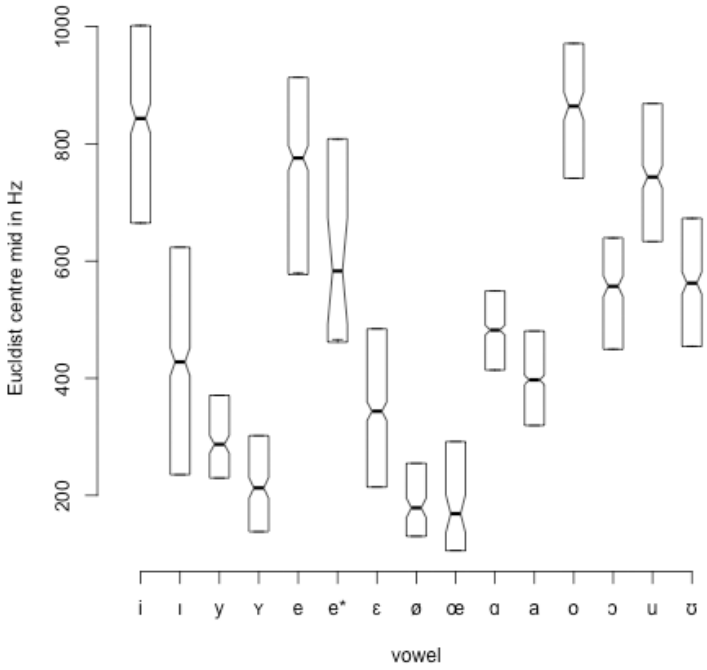


Figure 4: Boxplot of Euclidian distance in Hz for A- vs. B-vowel pairs for all stressed German vowels. *Kiel Corpus of Read Speech,* 26 female speakers.

16 See footnote 9.

The results in Figure 5 focus on the relations between the pairs /ɑ/–/a/, /e*/–/ɛ/, and /e/–/ɛ/, two of which have been claimed to exhibit a pure quantity contrast (cf. (1)). Our measurements show that all of these pairs exhibit the expected phonetic correlate, in accordance with their analysis as part of a single quality opposition on German.

The objection that at least some varieties of standard German might have a pure quantity contrast, at least for some oppositions, calls for a detailed study, focusing on the speech of maximally homogeneous groups or even individuals. This is because for phonetically similar sounds there is a danger that significant differences in the pronunciation of individuals become obscured by merging data. Even for a single speaker, systematic differences can be obscured by merging results pertaining to different segmental and prosodic contexts. The data in Figure 6 are based on the OLLO speech corpus, which contains minimally contrasting segment strings (cf. section 4). They demonstrate significant differences for the relative Euclidian distances for the /a/:/ɑ/ contrast compared in various segmental contexts, indicating for instance stronger contrasts in velar compared to labial contexts.

The ideal phonological corpora for establishing phonemic contrast are based on carefully controlled studies, where simplexes appear in identical carrier sentences and the speech of individuals can be examined separately.[17] In general, it holds that the demonstration of significant phonetic differences in a single context for a single speaker suffices to establish an active FAITH constraint in the phonological grammar of that individual.

Apart from demonstrating consistent phonetic correlates for phonological oppositions, there are various additional ways to test phonological analysis with speech corpora. The analysis predicts specific vowel qualities in the neutralization contexts, including only centralized vowels in unstressed closed syllables as in '*Gyr*[ɔ]*s* 'gyros' or only peripheral vowels in unstressed open syllables as in *B*[i]*kín*[i] 'bikini', [ɑ]*lásk*[ɑ] 'Alaska'. All subphonemic properties are predicted to conform to certain contextually determined restrictions such as only enhancement (rather than weakening) of gestures in strong prosodic positions (e.g. possible lengthening, never shortening, of vowels in stressed syllables). Subphonemic properties are further predicted to not exhibit paradigm uniformity effects (e.g. no difference in vowel length for the first vowel in *platónisch* 'Platonic' and *Platáne* 'plane tree', despite the presence of a long vowel in the base '*Pl*[ɑ:]*to* 'Plato'). At the same time, it is predicted that phonemic structure, including quality contrasts concerning the feature [±peripheral], can show paradigm uniformity effects (e.g. a peripheral unstressed vowel in plural '*Aut*[o]*s* 'cars', distinct

17 cf. the formant maps in Ramers 1988: 181ff

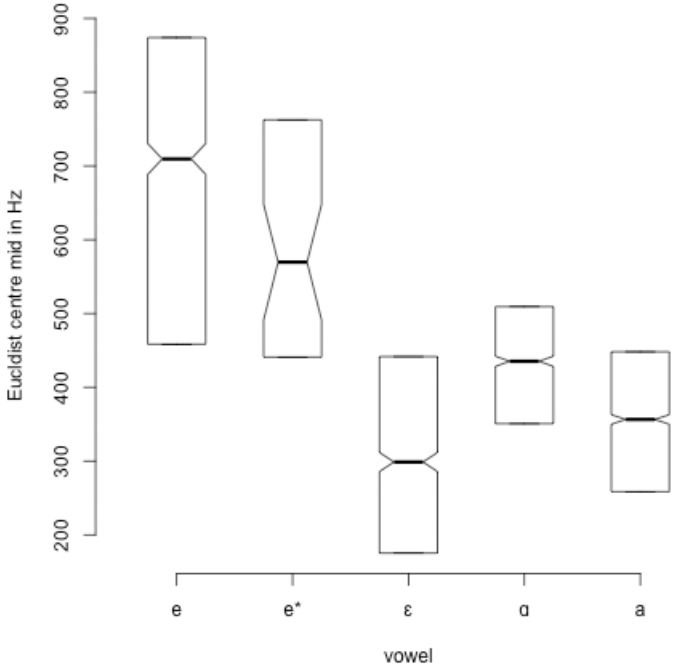


Figure 5: Boxplot of Euclidian distance in Hz for /e/, /e*/, /ɛ/, /ɑ/, /a/. *Kiel Corpus of Read Speech,* 26 female speakers.

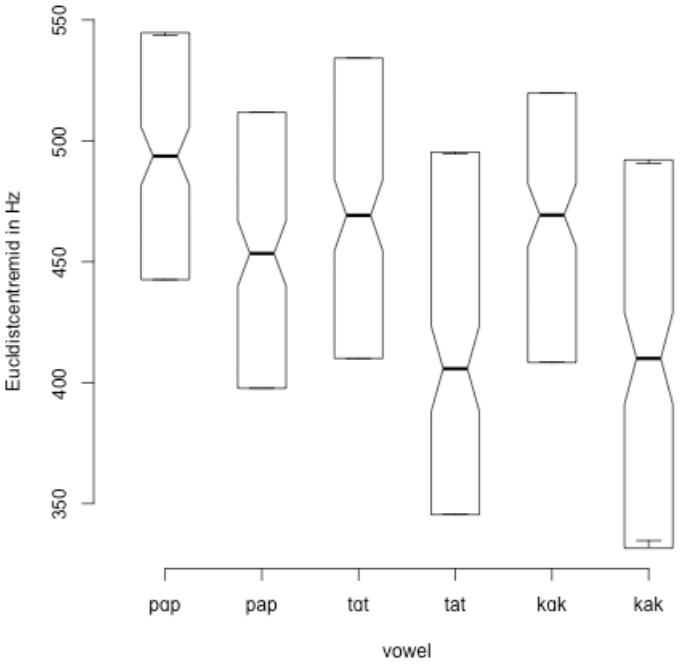


Figure 6: Boxplot of Euclidian distance in Hz for /a/ vs. /ɑ/ with different consonant contexts, each represented with 176–180 tokens. *OLLO corpus,* 5 Bavarian female speakers.

from the centralized vowel in '*Gyr*[ɔ]*s* 'gyros', to match the peripheral vowel in the singular '*Aut*[o] 'car'). For some preliminary studies of these types to verify the [± peripheral] opposition of German vowels, see Raffelsiefen (2016).

4 Data resources

Below we will briefly discuss the data resources used in our research: speech corpora, typological databases, electronically searchable word lists, and various word collections.

The corpora mentioned above, the *Kiel Corpus* (Kohler 1994) and *OLLO* (Wesker et al. 2005), have the advantage that they are provided with complete corrected segmental annotations but differ greatly in scope. The Kiel Corpus contains recordings of read connected speech, including 31,000 word tokens from 53 native speakers of German. *OLLO* contains recordings of read nonce words of the type CVC and VCV, presented in conventional German orthography (e.g. <pahp>, <papp>). It is based on 40 speakers divided into four separate regions and contains 2,700 recorded tokens per speaker. While confined to a subset of German phonemes, and arguably not containing German language material proper, the highly controlled environments yield valuable information about subtle contrasts, contextual influences, and regional differences.

A third corpus for German we frequently use is *Deutsch Heute* (Brinckmann et al. 2008), which includes recordings of roughly 1,000 words, including many loanwords, by 670 speakers covering all German-speaking areas. This corpus is well-suited to studying regional variation. It is, however, not suited to studying subtle contrasts as there are almost no minimal pairs and the words are read without carrier sentences. Moreover, some of the material is currently provided only with automatic segmental annotation, which needs to be corrected manually. We resort to special purpose-built corpora when necessary to study subtle phonological contrasts or specific paradigm uniformity effects.

Generally speaking, annotations cannot be assumed to be adequate, even when manually corrected. For instance, annotations for the *Kiel Corpus* mark all word-final full vowels as long, regardless of stress. As a result, a word like *Alaska* 'Alaska' has identical representations for the first two vowels, distinct from the last, which is transcribed as long (e.g. /Qal'aska:/, where Q = glottal stop). As was noted above, a study of neutralization patterns in German indicates a restriction to peripheral vowels (or schwa) in open syllables, in contrast to centralized vowels in closed syllables (i.e. /ɑ:'las.kɑ/). It goes without saying that proper annotations are a crucial prerequisite for meaningful phonological studies. (The reference to the *Kiel Corpus* in our measurements of the low vowels is restricted to stressed syllables for this reason.)

To establish markedness constraints, we consult typological databases such as UPSID (cf. section 2). At close sight, the results of such studies often raise questions. Consider again the case of markedness involving lip roundedness, which in fact involves two parameters, vertical lip compression and lip protrusion (cf. Ladefoged and Maddieson 1996: 295). These are implemented jointly in most languages, but what is the claim for each individual parameter? Worse problems arise with respect to the sort of phonetic length and quality differences observed in German. Basing a typological study on the results presented by Kohler (1992) or Eckert & Barry (2005), compared to the results proposed here, will greatly affect the outcome of typological work.[18] If for a relatively well-studied language like German there is so little consensus of how to present the basic vowel system then how does this bode for hundreds of less studied languages? Again, the central issue here is abstractness: comparisons are valid only if all studies subsumed in typological surveys conform to specific methods for conducting phonological analyses.

To study neutralization patterns, we use electronically searchable word lists including the CELEX databases for German and English (cf. Baayen et al. 1995) and pronunciation dictionaries (e.g. Wells (2000) for English, Krech et al. (2009) and Dudenband 6 (2015) for German). The CELEX databases have the advantage that they are searchable with regular expressions, allowing for the extraction of word lists matching specific patterns. These databases are useful for finding examples or getting a first impression concerning certain patterns. Their disadvantage is that they are far too small (ca. 50,000 entries for German CELEX), include no information on variation, and tend to exclude precisely the most valuable "marginal" words discussed above.

The pronunciation dictionaries are much more comprehensive (for instance roughly 150,000 entries in Krech et al. (2009)) and also include some useful information regarding variation (especially Wells (2000) and Dudenband 6 (2015)). However, they, too, contain relatively little information on the "marginal" words, especially acronyms. For foreign proper nouns they often list only the entirely unassimilated pronunciation pertaining to the source language (e.g. [prɔˈvɑ̃ːs] 'Provence' in Dudenband 6).[19] Electronic searches are tedious, as only specific grapheme strings can be submitted in search queries.

As was noted above, for the time being the perhaps most valuable data to establish constraint interaction consist of loan words and acronyms, even speech

18 Cf. Becker-Kristal (2010: 7ff) for an overview of different perspectives on vowel length in typological surveys.

19 These omissions are understandable given the main purpose of these dictionaries to provide information on the "correct" pronunciation, not least to meet their users demand to avoid possible social stigma.

errors, all of which involve a relation among an output and a given input form. Comparisons of these forms allow for systematic modifications of sound structure to be established, thereby providing a window on active constraints. The relevant adaptation patterns typically involve discrete decisions. Is French [bis'tʀo] 'bistro' borrowed into German by imitating peripheral [i] or by replacing it with a centralized [ɪ] (/bis'tʀo/ or /bɪs'tʀo/), by imitating the final stress or by shifting it to the initial syllable (/bɪs'tʀo/ or /'bɪstʀo/)? Does the pronunciation of the acronym GAL rhyme with /bal/ <Ball> 'ball' or with /vɑl/ <Wal> 'whale'? Discrete decisions of this type lend themselves to documentation in the form of transcriptions, as the choices of symbols can be assumed to be fairly reliable.[20] Unfortunately, the relevant data are nonetheless difficult to obtain, as even specialized dictionaries of abbreviations and acronyms (e.g. Steinhauer 2005), give no information regarding the pronunciation. As far as we know, there are currently no corpora for speech errors with phonological transcriptions or organized around phonological questions.[21]

The most valuable data to shed light on correspondence constraints also concern pairs of words, that is rhymes and paradigmatic alternations. Again, these data are often hard to find and, like loan word adaptation patterns and acronyms, ought to be backed up by experimental studies. The relevant collections will always pale in size compared to regular speech corpora but are likely to yield valuable insight into phonological systems. It is unclear how a strictly corpus-driven approach based on "raw speech" corpora alone could achieve this. In fact, the wider issue emerging from the above discussion of the problematic annotations in the *Kiel Corpus* is that proper annotation presupposes a thorough phonological analysis, to yield classifications of sounds which can be meaningfully compared.

References

Baayen, R. Harald, Richard Piepenbrock and Leon Gulikers. 1995. *CELEX2 LDC96L14*. Web Download. Philadelphia: Linguistic Data Consortium.

Becker, Thomas. 1998. *Das Vokalsystem der deutschen Standardsprache.* Frankfurt et al.: Peter Lang.

20 The outright rejection of transcriptions as "second-hand data", on the basis that it is impossible to verify their validity (Delais-Roussarie & Yoo 2014: 203), seems more justified when concerning subtle phonetic detail.

21 Here recordings would in fact be valuable to investigate the claim that only phonemic structure is affected by speech errors (Stampe 1973).

Becker-Kristal, Roy. 2010. *Acoustic typology of vowel inventories and Dispersion theory: Insights from a large cross-linguistic corpus.* Ph. D. Dissertation, Los Angeles: UCLA.

Beckman, Jill. 1998. *Positional Faithfulness.* Ph. D. dissertation. Amherst: UMass.

Boersma, Paul and David Weenink. 2016. *Praat: doing phonetics by computer* [Computer program]. Version 6.0.17 "http://www.praat.org/" (21 April 2016).

Brentano, Clemens. 1852. Romanzen vom Rosenkranz. Wikisource website. https://de.wikisource.org/wiki/Romanzen_vom_Rosenkranz (July 2017).

Brinckmann, Caren, Stefan Kleiner, Ralf Knöbl and Nina Berend. 2008. German Today: an areally extensive corpus of spoken Standard German. In *Proceedings 6th International Conference on Language Resources and Evaluation* (LREC 2008), Marrakesch, Marokko.

Chambers, John M., William S. Cleveland, Beat Kleiner and Paul A. Tukey. 1983. *Graphical Methods for Data Analysis.* Belmont, CA: Wadsworth & Brooks/ Cole.

Delais-Roussarie, Elisabeth and Hiyon Yoo. 2014. Corpus and Research in Phonetics and Phonology. In Jacques Durand, Ulrike Gut and Gjert Kristoffersen. *The Oxford Handbook of Corpus Phonology*, 193–213. Oxford: Oxford University Press.

[Dudenband 6] = Kleiner, Stefan, Ralf Knöbl and Max Mangold. 2015. *Duden – Band 6: Das Aussprachewörterbuch.* Berlin: Dudenverlag & Mannheim: IDS.

Eckert, Hartwig and William Barry. 2005. *The Phonetics and Phonology of English Pronunciation.* Trier: Wissenschaftlicher Verlag Trier.

Hixon, Thomas J., Gary Weismer and Jeanette D. Hoit. 2008. *Preclinical speech science: Anatomy, physiology, acoustics, perception.* San Diego, CA: Plural Publishing Inc.

Kohler, Klaus J. (ed.). 1992. Phonetisch-Akustische Datenbasis des Hochdeutschen. Kieler Arbeiten zu den PHONDAT-Projekten 1989–1992. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK),* 26.

Kohler, Klaus J. 1994. Lexica of the Kiel PHONDAT Corpus Read Speech. Volume I. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK),* 27.

Kohler, Klaus J. 1999. German. In: *Handbook of the International Phonetic Association.* Cambridge: Cambridge University Press.

Krech, Eva-Maria, Eberhard Stock, Ursula Hirschfeld and Lutz Christian Anders. 2009. *Deutsches Aussprachewörterbuch.* New York: De Gruyter.

Ladefoged, Peter and Ian Maddieson. 1996. *The sounds of the World's languages.* Oxford: Blackwell.

Leuninger, Helen. 1996. *Reden ist Schweigen, Silber ist Gold.* München: Deutscher Taschenbuch Verlag.

Lindau, Mona. 1978. Vowel Features. *Language* 54: 541–563.
Maddieson, Ian. 1984. *Patterns of sounds.* Cambridge: Cambridge University Press.
Maddieson, Ian and Kristin Precoda. 1990. Updating UPSID. *UCLA-Working Papers in Phonetics* 74: 104–114.
Peterson, Gordon and Harold Barney. 1952. Control Methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24: 174–185.
Prince, Alan and Paul Smolensky. 1993. Optimality Theory: Constraint Interaction in Generative Grammar. *Rutgers University Center for Cognitive Science Technical Report* 2, New Brunswick: Rutgers University Press.
Raffelsiefen, Renate. 2016. Allomorphy and the question of abstractness. Evidence from German. *Morphology* 26, Issue 3.
Ramers, Karl Heinz. 1988. *Vokalquantität und –qualität im Deutschen.* Tübingen: Niemeyer.
Stampe, David. 1973. *A Dissertation in Natural Phonology.* Ph.D. dissertation, University of Chicago, published 1979 by Garland Press, New York.
Steinhauer, Anja. 2005. *Duden. Das Wörterbuch der Abkürzungen.* 5th revised ed. Mannheim et al.: Dudenverlag.
Vennemann, Theo. 1991. Skizze der deutschen Wortprosodie. *Zeitschrift für Sprachwissenschaft* 10: 86–111.
Wängler, Hans-Heinrich. 1964. *Atlas deutscher Sprachlaute.* 3rd revised ed. Berlin: Akademie Verlag.
Wells, John. 2000. *Longman Pronunciation Dictionary.* 2nd edition, Harlow: Pearson Education Limited.
Wesker, Thorsten, Bernd Meyer, Kirsten Wagener, Jörn Anemüller, Alfred Mertins and Birger Kollmeier. 2005. *Oldenburg Logatome Speech Corpus (OLLO) for Speech Recognition Experiments with Humans and Machines*, Proc. of Interspeech, 4–8. Lisbon, Portugal.

Don Tuggener, Martin Businger

Needles in Haystacks: Semi-Automatic Identification of Regional Grammatical Variation in Standard German

Abstract This paper lays out a semi-automatic approach to identifying regional variation in the grammar of Standard German. Our approach takes as input manually defined templates of grammatical constructions that are automatically instantiated over a corpus collected from regional newspapers. These instantiations are automatically ranked by a metric that quantifies how specific an instantiation is for a region. Ranked lists of instantiations are compiled that contain instantiations specific to a region and are scanned manually by linguists to identify those that denote grammatical variants of Standard German. This approach enabled us to discover variants that so far have not been documented. With respect to research on variation within standard languages as seen from a more general perspective, we aim to contribute towards research strategies that clearly rely on empiricism rather than on intuition or bias.[1]

Keywords Association measures, corpus-driven approaches, diatopic variation, grammatical variation, standard language

1 Introduction

Varieties of a language can display differences in usage at any linguistic level, e.g. pronunciation, grammar, vocabulary or spelling. Variation regarding a feature of one of these linguistic levels—an intralinguistic feature—can correlate with extralinguistic factors, i.e. diastratic, diachronic, diaphasic or diatopic factors.

1 This paper received the support of the *Swiss National Science Foundation (SNSF)* and of the *Austrian Science Fund (FWF)*; grant numbers: *SNSF* 100015L_156613; *FWF* I 2067-G23. We would like to thank Gerard Adarve, Nicole Zellweger, Regula Gass, Reinhard Kunz, Marek Konopka and an anonymous reviewer for their help or comments on earlier versions of this paper.

This paper focuses on the correlation between grammatical variation and the diatopic dimension. Nevertheless, the approach and the methods laid out below are, in principle, applicable to any linguistic variation phenomena that correlate with features pertaining to any extralinguistic dimension.

This work is part of the project *Variantengrammatik des Standarddeutschen* ("Regional Variation in the Grammar of Standard German", cf. http://variantengrammatik.net/) which aims to identify and document grammatical variation in Standard German based on a regionally balanced corpus. For a detailed description of the project design, see Dürscheid and Elspaß (2015). We advocate an approach where language norms constituting a standard language—Standard German in our case—are to be reconstructed based on actual language usage; see Elspaß and Dürscheid (2017) for an extensive discussion on the term *Gebrauchsstandard*, i.e. 'standard language as it is used', and its interpretation in the context of the research project. The project will primarily result in an open-access website that compiles the project's findings and that serves as a searchable database of grammatical variation of Standard German (Dürscheid et al. in prep.).

The corpus compiled for this research project consists of texts from 68 online newspapers that were crawled for approximately one year, thus representing the German *Gebrauchsstandard* from all countries of Europe where German is used as an official language, divided into 15 regions (see Figure 1) based on the "Variantenwörterbuch" (first edition 2004 [= Ammon et al. 2004] and second edition 2016 [= Ammon/Bickel/Lenz et al. 2016], see e.g. map for Germany on p. LIII). The corpus contains roughly half a billion words distributed over 1.5 million articles which have been automatically processed with computational linguistics software (most importantly lemmatization, part-of-speech tagging, morphology, and dependency parsing). This corpus constitutes the basis for our experiments.

Clearly, reading a large text corpus like ours to discover regional grammatical variants is cumbersome and infeasible. Thus, the appeal of (semi-)automated methods that promise to alleviate much of the work is strong. A key interest of this contribution is thus to determine how well automatic and statistical methods from corpus and computational linguistics can assist grammarians in identifying regional grammatical variants. We propose a processing pipeline in which expert linguists and automatic ranking algorithms work together and evaluate how fruitful this collaboration is (Figure 1).

We proceed as follows. In section 2, our semi-automatic approach to identifying regional grammatical variants is described in detail and is compared to related work. In section 3, we examine selected examples of the results and discuss them in the context of recent research on grammatical variation within Standard German. The paper concludes with a summary (section 4).

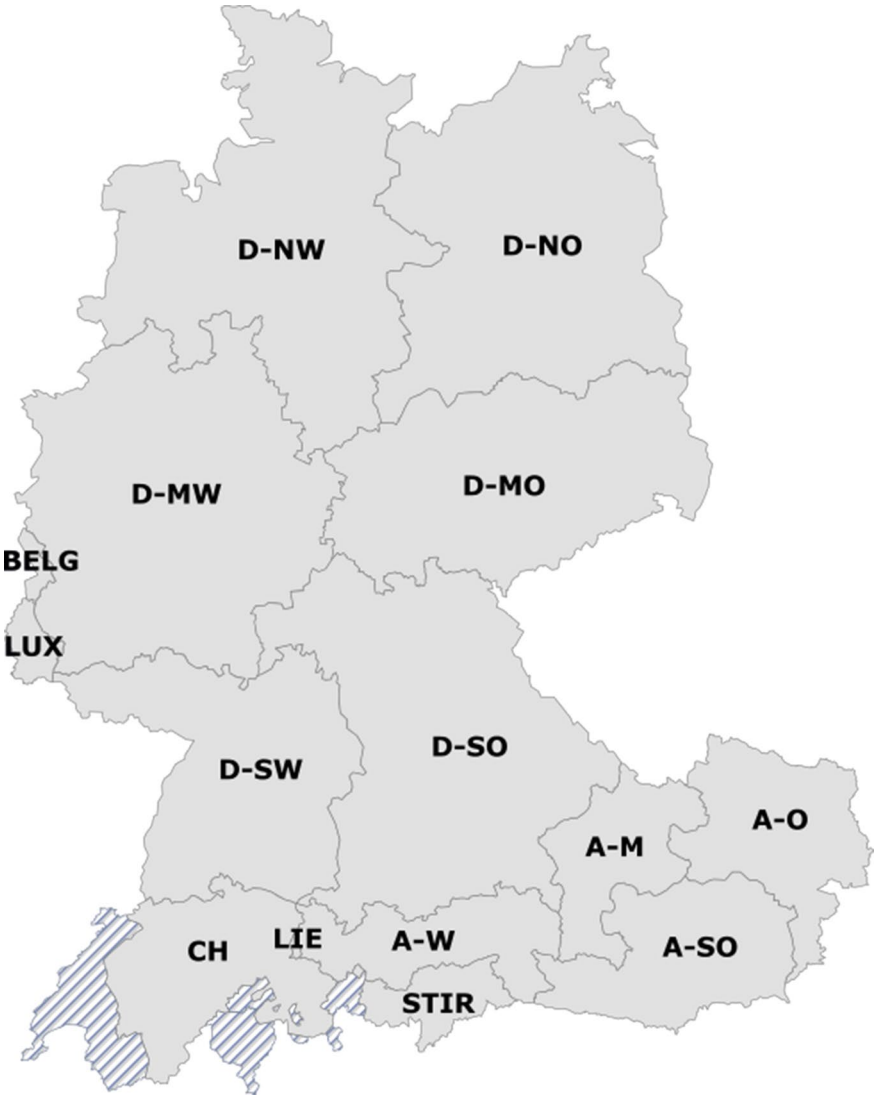


Figure 1: European countries and regions with German as an official language, with subregions.

2 Semi-automatic identification of grammatical variants

Before turning to our own approach (section 2.2), we briefly discuss relevant related work (section 2.1). Sections 2.3 and 2.4 explain our choice of a suitable ranking metric in detail.

2.1 Related work

One way of discovering grammatical variants is to have speakers from one country or region read newspapers of another country/region and mark the constructions that strike them as 'odd'. These constructions are then queried in a corpus and their distributions are analyzed statistically to verify whether there is sufficient support to categorize them as variants. Obviously, this approach is time-consuming and expensive. Another approach is to gather variants previously described in the literature and then query those in a corpus. The obvious drawback of this method is that it does not allow for any new variants to be identified.

The natural appeal of a corpus-driven approach therefore is its ability to overcome the drawbacks of the two methods described above. Firstly, it requires less time for a machine to read through large corpora, and secondly, the machine does not rely (heavily) on a priori assumptions about variation. Clearly, analyzing all random combinations and permutations of lexeme sequences and their various linguistic properties is infeasible even for smaller corpora. Furthermore, one cannot expect all grammatical constructions to show regional variants—on the contrary: we expect most constructions to be distributed homogeneously. Hence, using some initial and loose linguistic intuitions about which phenomena can be expected to show regional variation is a reasonable approach to help reduce search space.

Our work aligns with corpus linguistic research that aims to compare genres, registers, or varieties of languages. One area therein is the comparison of second language learner corpora to native speaker corpora, e.g. Laufer and Waldmann (2011), Cao and Xiao (2013), and Yoon (2016). Another area evolves around grammatically distinguishing the varieties of e.g. English, e.g. Mukherjee and Hoffmann (2006), Mukherjee (2009), and Xiao (2009). In this area, our approach is most closely related to Schneider and Zipp (2013), who also used an automatic dependency parser in their approach. An important advantage of using a dependency parser over so-called 'window-based' methods is that dependency parsing can tackle long-distance dependencies between lexemes that fall out of the window size. Window-based methods slide a window of a predefined size (e.g. two or five consecutive words) over the sentences in the corpus and analyze

the distribution of re-occurring word sequences. We experimented with different window-based approaches, including complex ngrams (i.e. replacing certain lexemes with their part-of-speech tags) along the lines of Bubenhofer (2015), but struggled to find a setup that yielded ranked lists which contained regional grammatical variants.

The aim of Schneider and Zipp (2013) was to identify novel combinations of verb and preposition in Indian and Fiji English in the International Corpus of English. They compared a fully manual approach to a semi-manual one. In the fully manual approach, the researcher first queried, on the one hand, a list of prepositions known to be productive and, on the other hand, an additional two prepositions that are commonly assumed to show variation in the literature. The combinations found were then compared to dictionaries that contain known variants, and those not contained in the dictionaries were labeled as unrecorded. The semi-automatic approach used a dependency parser and a metric to rank all found verb-preposition combinations which were then evaluated by the linguist. To automatically obtain ranked lists of verb-prepositions combinations, they scored each lexicalized combination in the Fiji and Indian English subcorpora with an observed over expected count ratio (compared to the BNC corpus). Combinations that were considered "unexpected" by the ratio were ranked high and then manually evaluated by a linguist.

The fully manual approach has the advantage of being highly accurate, i.e. the linguist will only pick those query results which are indeed variants. Clearly, the drawback of this method is that it is time-consuming and requires the researcher to know beforehand which lexical items (in their case a set of prepositions) are assumed to induce variation. The semi-manual, parser-assisted approach, on the other hand, has the advantage of not requiring a priori assumptions about the variation of specific lexical items but proceeds in a theory-agnostic, purely corpus-driven fashion. Its drawback is that automatic parsing yields errors and thus reduces the precision of the approach (returning false positives and missing true positives due to parsing errors).

In contrast to Schneider and Zipp (2013), we do not solely focus on combinations of verbs and prepositions. We are interested in all aspects of verbs and their subcategorization frames. That is, we query verb lemmas and all grammatical functions that they subcategorize for (e.g. direct/indirect objects, prepositional phrases, subclauses etc.). Furthermore, we are interested in word formation phenomena, e.g. the combination of verb stems and prefixes, and whether there are regional preferences for certain combinations. Another important difference of our setting to that of Schneider and Zipp (2013) is that our corpus comprises 15 subcorpora (corresponding to geographical regions), rather than two or three. Hence, computing the Observed-Expected ratio used in Schneider and Zipp (2013) would be computationally expensive, since it requires counting each verb

and preposition both together and separately for each subcorpus and the concatenation of the remaining subcorpora to decide whether a combination of a verb and a preposition is "unexpected". Our ranking metric requires less counting and does not need to partition the subcorpora in a one-versus-the-rest fashion to calculate a score for the specificity of a construction in a certain region.

2.2 Pipeline approach

Accounting for the discussion above, we define the following semi-automatic pipeline to discover novel grammatical variants:

Table 1: Pipeline approach.

1	Identify a general grammatical pattern that is assumed to show regional variation, e.g. verb valency.	Manual
2	Translate the pattern to a path or template construction in the dependency trees annotated in the corpus.	Manual
3	Instantiate the template over the corpus, track counts per region.	Automatic
4	Analyze the distribution of each instantiation with respect to the regions. Return a list of instantiations ranked by their specificity for a particular region.	Automatic
5	Inspect the list and manually distinguish between grammatical, orthographic, and lexical variants (and noise).	Manual

To illustrate the process, we walk through the following example: In step 1, we assume that verbs show regional variants with regard to the preposition that they subcategorize for. We formulate the template: verb + preposition (step 2), i.e. only the part of speech of the two items as well as their dependency relation (the preposition is governed by the verb) are specified. Next, in step 3, we automatically extract all lexicalized instantiations of the template from the dependency trees in the corpus, counting their occurrence per region. The following sentence is an example of an instantiation:

(1) Zunächst setzte sich Borna über Turbine Leipzig durch […][2]
first VERB REFL *Borna* over *Turbine Leipzig* VERB-PREFIX
'First, *Borna* won against *Turbine Leipzig* […]'

2 http://www.lvz.de/Region/Borna/Zwei-Heimsiege-Aufstieg-und-Belohnungsspiel (10 February 2017).

Given the automatic dependency analysis shown in Figure 2, we extract the following instance tuple:[3] <durchsetzen, über, D-Nordwest, 1> (i.e. <verb, preposition, region, count>).

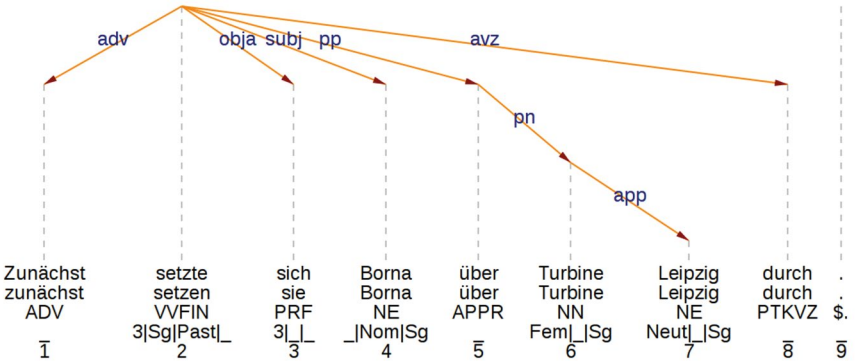


Figure 2: Output of the dependency parser for example sentence 1.

Having traversed all dependency trees in the corpus, the instantiations found are analysed and ranked with respect to their specificity for a region using a metric (cf. section 2.3) in step 4. Our example instantiation from above will rank high in this list because the verb *durchsetzen* 'prevail' commonly subcategorizes for the preposition *gegen* 'against' instead of über 'over' (see discussion of this verb below in section 3.3). Instantiations like *legen+in* ('put+in') will have a low rank, since they occur frequently in all regions.

In step 5, this list is inspected by a linguist to cherry-pick the instantiations that denote grammatical variants. This is necessary because the metric ranks all 'peculiar' constructions high, which means that orthographic (e.g. *ss* instead of *ß* and vice versa) and lexical (e.g. *paraphieren* 'to initial') as well as noise (e.g. verb instances containing encoding errors of *Umlauts*) are ranked high because the metric is not able to distinguish them.

Having outlined the approach, we turn to its core next, the ranking metric.

3 Note that in order to get the correct verb lemma (*durchsetzen*), we have to attach the separable verb prefix (*durch*) to the stem (*setzen*). Otherwise, the instantiation would wrongly be attested to the verb *setzen*. Fortunately, the dependency parser reliably identifies separated verb prefixes.

2.3 Ranking metric

The corpus linguistic literature contains a vast variety of metrics that aim to identify linguistic items that manifest some desirable properties (association, heterogeneous distribution etc.). Providing a comprehensive overview is beyond the scope of this work, and we refer readers to e.g. Evert (2004) and Gries (2008). Instead, we outline the requirements for a metric in our setting and motivate our choice based on them.

In our setting, the task of the metric is to assign a high rank to grammatical variants that occur in a (limited) set of regions. Hence, one criterion for the metric is that a template instantiation should be ranked high if it only occurs in a small number of regions. In other words, the rank of an instantiation should increase with the decreasing number of regions that contain it. Among those instantiations with such limited coverage in the corpus with respect to the regions, we want those to rank high that have a high frequency. We favor high frequency instantiations because we want to avoid the problem of defining an arbitrary minimum frequency threshold for including phenomena in the variation grammar wherever possible. Low-frequency instantiations also often cause problems with low expected values in subsequent statistical analyses (e.g. Chi Square). In addition, favoring high frequency phenomena acts as a natural filter against occasionalisms, typing errors and the like as well as various preprocessing problems, such as encoding errors and faulty dependency parses, which is essential since we work with real-world data and automatic preprocessing.

One metric that perfectly combines both desiderata is *Term Frequency Inverse Document Frequency* (TF IDF), well-known in Information Retrieval. TF IDF is widely used, e.g. for document indexing for search engines. A term is regarded as highly indicative for a document if it occurs frequently in the document (term frequency; TF), but at the same time occurs only in a small number of other documents in a collection (inverse document frequency; IDF). In our setting, we treat the template instantiations as the terms, and the regions as the documents.

More specifically, we calculate the normalized TF of a template instantiation t_i given a region r_j (e.g. <verb, preposition, region, count> = <durchsetzen, über, D-Nordwest, 216>) as:

$$TF = \frac{count(t_i, r_j)}{\sum_{k=1}^{n} count(t_k, r_j)}$$

i.e. by dividing the count of t_i in region r_j by the sum of all counts of all instantiations in r_j. This division normalizes TF to the size of the subcorpus r_j and lets us compare subcorpora of different sizes.

IDF is simply (the logarithm of) the ratio of all regions and the regions r that contain the template instantiation t_i:

$$IDF = \log_2 \frac{count(r_{k...n})}{count(r_{k...n}) \ni t_i}$$

TF IDF is the product of the two, i.e.:

$$TFIDF = TF \times IDF$$

Using this approach, we are able to rank all template instantiations both per region and for all regions combined by creating corresponding ranked lists (one for each region and one for all regions combined).

TF IDF has the advantage that it is relatively cheap to compute compared to other metrics like Observed-Expected ratios or Mutual Information because it does not require access to the counts of the individual components in the constructions (e.g. the separate counts of a verb and a preposition in the subcorpora, which are required by Mutual Information to calculate their association strength).

However, one downside of TF IDF is that in the IDF calculation, the dispersion of an instantiation (i.e. t_i) is not taken into account. This means that looking up the number of regions that contain t_i does not account for how well t_i is supported in those regions. For example, t_i might only occur once in a comparably large subcorpus, but with high frequency in three smaller subcorpora. However, all these occurrences are weighted equally. Conversely, another template instantiation t_k might occur frequently in the larger subcorpus and only once in each of the three smaller subcorpora. For both t_i and t_k, the IDF value will be the same, since they occur in an equal number of regions. However, their dispersions or distributions in the subcorpora are vastly different, and we would like our metric to reflect that. Thus, we introduce a notion of dispersion to the TF IDF calculation by multiplaying it with the *DISP* parameter, which is based on the count distributions, more specifically their residuals, and calculated as follows:

$$residual(t_I, r_j) = \frac{observed(t_i, r_j) - expected(t_i, r_j)}{\sqrt{expected(t_i, r_j)}}$$

$$DISP(t_i, r_j) = residual(t_i, r_j) - mean(residuals(t_i, r_{j...n}))$$

That is, we subtract the mean of all of t_i's residuals from that of the current region r_j. Note that if t_i's residual in r_j is above the mean, this yields a positive number and vice versa. Hence, all template instantiations whose residual in a given region is below the mean of all its residuals will render the TF IDF score

negative for that region and will rank it low in the list of specific constructions. Conversely, all instantiations with a positive difference to the residuals' mean will get a boost in the ranking. Our final metric then simply consists of:

$$TFIDFDISP = TF \times IDF \times DISP$$

There are other noteworthy metrics that rank construction in relation to the heterogeneity of dispersion. A whole family of statistical tests can serve as such a metric, e.g. Chi Square. One common problem of these tests (which we also encountered during preliminary experiments) is that they tend to yield high significance levels for low-frequency phenomena in large corpora (Gries 2008). Since we are interested in highly frequent phenomena, this is a clear disadvantage. The same applies to (Pointwise) Mutual Information-based metrics. An interesting, intuitive and easily computed metric of dispersion is presented in Gries (2008), called *deviation of proportions*. It is also based on normalized values for observed and expected frequencies and their differences, similar to our *DISP* parameter. We will empirically compare our metric to the unaltered version of TF IDF and Gries' *deviation of proportions* (Gries DP henceforth) in the next section.

2.4 Comparison of metrics

In this section, we compare the ranked lists that emerge when we apply the three ranking metrics outlined above, i.e. TF IDF, TF IDF DISP, and Gries DP to a set of instantiated templates. The instantiations that we rank stem from the combination of two verb-related templates, i.e. verbs and the (lexicalized) prepositions they subcategorize for,[4] and verbs and the (unlexicalized) grammatical functions in their subcategorization frame.[5] We compare the lists by assigning the top 100 instantiations in each to five categories: grammatical variants (which we are interested in), lexical variants (interesting, but not in our focus), *ss/ß* alternation (irrelevant in our case), non-variants (instantiations that are overrepresented in some area of the corpus due to the sampling process, e.g. *sich qualifizieren* 'to qualify' with reflexive morpheme *sich* is ranked high because of oversampling of the sports section), and preprocessing/encoding errors (noise in the corpus). The distribution of the instantiations over these categories can then serve as an estimate of how fruitful it is for a researcher to manually scan each list in terms of the number of returned novel variants, which serves as an evaluation.

4 An example instantiation is: *ersuchen + um* 'to request sth.'.

5 E.g.: *beantragen* 'to request, to apply for' + *dative object* or *beantragen + accusative object.*

As mentioned above, we are looking for phenomena that feature a solid support in the corpus and are thus interested in high frequency instantiations. To evaluate how well the metrics perform in this regard, we count how many of the top 100 instantiations in each list have a frequency of at least 10 occurrences. Note that for Gries DP, all counts in the corpus are considered, while for the TF IDF metrics only the counts in the respective region where an instantiation was ranked high are taken into account (thus the overall occurrences are even higher). To our surprise, we found that in the Gries DP list, only 1 of the top 100 instantiations has a corpus frequency of at least 10, while the top 100 lists created by TF IDF and TF IDF DISP feature 81 and 80 instantiations respectively, with a frequency over 10 in the region where they were ranked high. The Gries DP metric seems to suffer from oversensitivity to low count phenomena, at least in our setting.[6] Since we deem instantiations with a count below 10 as not sufficiently supported in the corpus, we removed all instantiations with a frequency below 10 from the Gries DP list, and then again took the top ranked 100 among the remaining instances for the further comparison.

Next, we analyze the top 100 ranked instantiations of the categories introduced above.

Table 2: Category breakdown per metric.

	Gries DP	TF IDF	TF IDF DISP
Preprocessing / encoding errors	19	31	32
ss/ß alternation: *begrüssen, begrüßen* 'to welcome'	28	43	27
Lexical variants: *paraphieren* 'to initial'	27	11	21
Non-variants	11	8	7
Grammatical variants	15	7	13

As shown in table 2, the filtered Gries DP list and the TF IDF DISP list return 13 to 15 instantiations that denote grammatical variants, while the TF IDF list only contains 7. TF IDF also returns the most *ss/ß* alternations (43), which the added DISP parameter is able to reduce (to 27). Gries DP is most robust against ranking preprocessing and encoding errors, but returns more lexical and non-variants than TF IDF DISP.

An interesting question is whether the different metrics return an overlapping set of instantiations in their top 100 lists or whether they favor different

6 An issue in the calculation of Gries DP in this respect is that it takes the absolute value of the differences between observed and expected values. Low count instances with a high negative difference to the expected value (which are based on normalized subcorpora sizes) therefore drastically increase the sum of the differences. Furthermore, the metric does not take into account the overall frequency of an instance, unlike TF IDF.

instantiations. We measure the overlap of the instantiations in each list in a pairwise manner in table 3.

Table 3: Pairwise overlap in the ranked lists.

Gries DP ∩ TF IDF	19
Gries DP ∩ TF IDF DISP	20
TF IDF DISP ∩ TF IDF	73

Clearly the ranked lists of the TF IDF metrics are more similar to each other than to the Gries DP list. Yet more than 25% of the instantiations in their lists are unique. Compared to the Gries DP list, there is little overlap with the TF IDF metrics. This suggests that the two approaches are complementary. Indeed, if we combine all the grammatical variants found in the three top 100 lists, we obtain a total of 22 unique grammatical variants.

One aspect that distinguishes the variants found in the Gries DP list and the TF IDF lists is their average frequency in the corpus compared to the average frequencies of the variants in the respective regions where the TF IDF metrics found them, as shown in table 4.

Table 4: Average frequency of variants found per metric.

	# Variants	Avg. frequency
Gries DP	15	42 (whole corpus)
TF IDF (region)	7	77 (region)
TF IDF DISP (region)	13	138 (region)

The table shows that the variants found in the Gries DP list have a much lower frequency compared to the TF IDF based variants. Furthermore, half of the 15 variants in the Gries DP list have a frequency below 15. Given a corpus of over half a billion tokens, the question arises whether such counts provide enough support to claim a variant.

Another downside of Gries DP is that it does not indicate directly which subcorpora (in our case regions) drive a high deviation of proportions,[7] if one is found, while the TF IDF-based measures can return ranked lists for any partition of the subcorpora or the whole corpus. Hence, based on the TF IDF measures, we can easily investigate instantiations that are specific to a given region or country.

After the comparison of the metrics, we now turn to some examples of newly discovered grammatical variants.

7 One could look at high positive differences between observed and expected, though.

3 Result examples: unknown grammatical variants

This section aims to illustrate the potential of the method by focusing on a small selection of results. After some initial remarks on the state of research and an overview of the results, we turn to specific examples from the areas of word formation and valency that we found using our approach.

3.1 Grammatical variation at different linguistic levels

Grammatical variation phenomena can be assigned to either morphology or syntax. In the field of morphology, we find areal (regional) variation in terms of both word formation and inflection. A vast array of morphological variants has been documented in the first and second edition of the *Variantenwörterbuch* (Ammon et al. 2004 and Ammon/Bickel/Lenz et al. 2016 respectively), which is undoubtedly the most comprehensive reference work on linguistic variation in the (written) German standard language to date. The *Variantenwörterbuch* aims primarily to document lexical variation, but it also includes variation phenomena in inflection (e.g. plural forms of nouns) and in word formation. As for syntax, the *Variantenwörterbuch* documents some variation with regard to valency, but syntactic phenomena are not taken into account systematically. This reflects the fact that research on variation within Standard German has traditionally focused on the lexicon and on morphology, rather than on syntax (cf. Niehaus 2015).

The semi-automatic approach outlined above is inherently not restricted to 'one-word-phenomena'. It has proven to be successful with a range of corpus findings in relation to word formation and valency (subcategorization). Overall, besides reproducing 23 previously known variants (i.e. documented in the *Variantenwörterbuch* or in at least one other relevant reference work for Standard German grammar, cf. examples below), we were able to discover 30 previously undocumented variants. In the next section, we present examples of areal grammatical variation still undocumented in relevant reference works. These phenomena were detected by using the pipeline approach described in section 2.

3.2 Word formation

The reflexive verbs s*ich berappeln* and *sich aufrappeln* both mean 'to stand up again' and, in a more figurative sense, 'to pull oneself together'. The key difference between the two verbs is a morphological one: while the verb *berappeln* has the unstressed, inseparable prefix *be*, the verb *aufrappeln* has the stressed and separable prefix *auf*. As is shown on the map in Figure 3, s*ich berappeln* is

found only in newspapers in Germany (and, occasionally, Belgium and Luxembourg). It is not attested in corpus texts from Austria or Switzerland. Note that *sich berappeln* is not mentioned as a regional variant in the *Variantenwörterbuch* (neither in Ammon et al. 2004 nor in Ammon/Bickel/Lenz et al. 2016). Neither does duden.de,[8] among the most widely used online works of reference, mention any regional restrictions on the use of *sich berappeln*. One might wonder if this 'gap' is purely accidental or can be attributed to a larger fundamental factor. We argue for the latter in the following section.

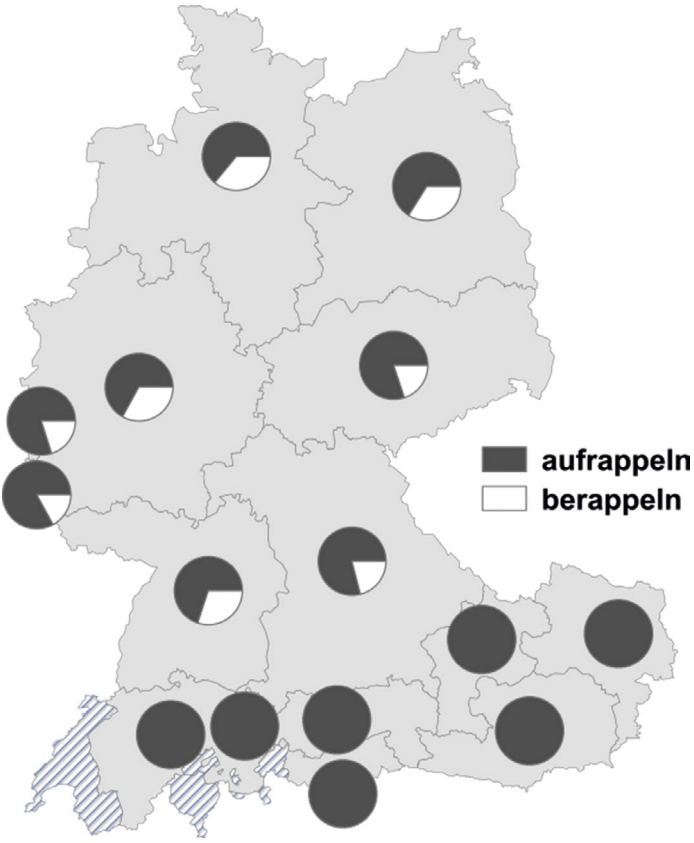


Figure 3: Distribution of *sich berappeln* vs. *sich aufrappeln.*

If only the—traditionally prevailing—manual method is used, linguistic features of Standard German that are used exclusively or mainly in Germany tend to pass unnoticed as regional variants by linguists (cf. Dürscheid and Sutter

8 duden.de (9 February 2017).

2014). This is due to a widespread bias in which the Standard German language of (Northern) Germany is thought to define the (only) norm (Schmidlin 2011: 208). According to Clyne (2004: 297), the varieties of pluricentric languages like German usually relate asymmetrically, with one variety dominating. Characteristic of such situations is the following, among other things: the dominant (D) variety has more effective political and economic resources for being exported, e.g. by means of reference works (dictionaries, textbooks etc.); users of the D variety may believe that there is no linguistic variation in *written standard* language; users of the D variety, as far as they notice differences between their own D variety and another variety, consider such other varieties as "exotic, cute" and, most importantly, "non-standard" (Clyne 2004: 297). This attitude is the basis of what can be identified as 'ideology of homogenism' (Elspaß and Niehaus 2014).

German as used in Germany clearly plays the role of the D variety. As a result, Germany-specific variants are less frequently marked as national or regional variants in reference works than e.g. national variants as found in Austria. This has been shown by systematic research on numerous grammar reference works (see Dürscheid and Sutter 2014 for details).

In this context, a second example worth noting is *bepöbeln* in contrast to *anpöbeln* 'to accost, to verbally abuse'. In our corpus, *bepöbeln* is confirmed to be used exclusively in Germany (in all regions except D-southwest; mainly in D-northwest). Again, *bepöbeln*, like *berappeln*, is not mentioned in the *Variantenwörterbuch* (either edition).

To conclude, the two examples, *sich berappeln* and *bepöbeln* indicate that a (semi-) automatic, at least partially corpus-driven, and thus less biased approach is superior to a purely manual one when it comes to identifying linguistic features of the dominant variety of a pluricentric language.

In the next section, we turn to examples of variation in subcategorization frames of verbs.

3.3 Valency

As a first example on valency, let us turn to the reflexive verb *sich durchsetzen* 'to prevail (against)', which can be combined with more than one preposition without difference in meaning (but note the caveat in footnote 10): *gegen,* über and *gegenüber* (meaning 'against'). *Gegen* is, as expected, by far the most frequently used preposition with *sich durchsetzen* in the corpus (black on the map in Figure 4). In contrast, the preposition *über* (gray on the map)—the one preposition that ranked high in combination with *sich durchsetzen* in our metric—is used almost exclusively in the center-east of Germany (one of the six predefined

German subregions) (cf. example (1) in section 2.2). A third attested preposition is *gegenüber*, which is generally rare and not restricted to particular regions (see Figure 4).

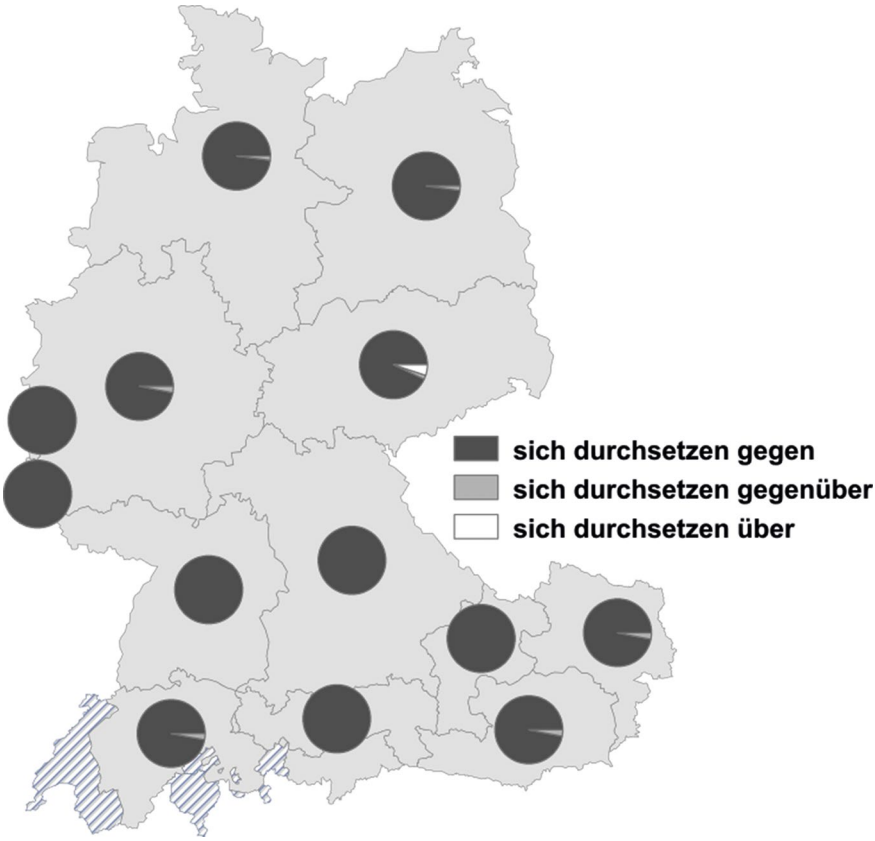


Figure 4: Distribution of *sich durchsetzen gegen* vs. *sich durchsetzen gegenüber* vs. *sich durchsetzen über.*

The verb *durchsetzen* is not listed in the *Variantenwörterbuch* (first and second edition) and it is therefore not possible to find any reference to prepositions selected by this verb there. The *Wörterbuch der Präpositionen* (Müller 2013)[9] has the prepositions *gegen* and *gegenüber* for *durchsetzen*, but not *über*—the one preposition that is of interest here because of its diatopically restricted usage.

9 This dictionary does not consider regional variation, but lists a large number of German verbs, adjectives and nouns with their respective prepositions.

We conclude that it is a hitherto unknown fact that *sich durchsetzen* is used with the preposition *über* in Standard German texts.[10]

A second example of regional variation in subcategorization frames is the verb *verlautbaren* 'to announce (officially), to proclaim'. According to the instantiations found in the corpus, *verlautbaren* ranked high in terms of our metric when governing a direct object NP.

The manual analysis of the phenomenon (in and after step 5, cf. section 2.2) proved to be complex. It is necessary to distinguish between several formal types of objects:

- (A) Nominal and pronominal objects: use of indefinite pronouns like *nichts* 'nothing' or *etwas* 'something' can be confirmed in almost all countries/regions without regional preferences. Examples with objects in the form of indefinite pronouns were therefore excluded (and are not represented on the map in Figure 5). Instead, only examples with a 'full NP'[11] object (including examples with full NP subjects in passive sentences as (2a) below) were counted.
- (B) Object clauses: subordinate clauses introduced by the subjunction *dass* 'that' or object clauses without subjunction (see example (2b)) together constitute one category.
- (C) No object (intransitive): usages of *verlautbaren* without any object at all commonly appear in a subordinate clause headed by *wie* 'as' which depends on the matrix clause (see example (2c), where the matrix clause is left out).

(2) a. Erst am Samstag soll […] das Endergebnis verlautbart werden.[12]
Only on Saturday is-said the final-result announced PASSIVE-AUX
'The result will not be announced until Saturday.'

b. Das Auswärtige Amt verlautbarte, die Echtheit des Videos
The *Auswärtige Amt* announced the authenticity of-the video
werde noch geprüft.[13]
PASSIVE-AUX still verified
'The Federal Foreign Office [of Germany] announced that the authenticity of the video remains to be verified.'

10 It must be noted that 35 out of 36 manually inspected corpus examples of *sich durchsetzen über* (= 97 %) were found in the sports section of the respective online newspapers. No such preference for a specific text type can be observed for *sich durchsetzen* when governing one of the other prepositions (*gegenüber* or *gegen*). Further research as to the (non-) interchangeability of the three prepositions governed by *sich durchsetzen* is necessary.

11 By the informal term "full NP", we refer to a nominal phrase headed by a noun, not a pronoun.

12 http://derstandard.at/1350260818406/Wahlergebnis-fruehestens-am-Samstag (10 February 2017).

13 http://www.schwaebische.de/region_artikel,-Filiz-G-soll-angeblich-freigepresst-werden-_arid,5227115_toid,351.html (22 March 2012).

c. Wie am Wochenende verlautbart wurde, […][14]
As on-the weekend announced PASSIVE-AUX
'As was announced on the weekend, […]'

In the resulting map (Figure 5), *verlautbaren* governing a full noun phrase (i.e. excluding pronouns) functioning as the object (black on the map; cf. example 2a) is contrasted with examples where the verb governs a clausal object or no object at all (white; cf. example 2b/c).

To sum up: in the Austrian regions, examples with full NP-objects constitute between 17 % (A-southeast) and 35 % (A-west). By contrast, in the middle and northern regions of Germany, this phenomenon is rare.

It is therefore possible to surmise that intricate and 'non-intuitive' variation phenomena, like the case of *verlautbaren,* would probably not be detected with a purely manual approach.

Let us now turn to a third example. In the area of verb valency, the diatopically conditioned alternation between reflexive and non-reflexive usage of certain verbs has received some attention in the literature. It has been presumed that speakers and writers of German in Austria tend to often use the reflexive pronoun *sich* with several verbs (Ebner 2008: 44f., Ziegler 2010). Current research has confirmed the alleged tendency to some extent (Dürscheid et al. in prep.). For example, the verb *erwarten* 'to expect' can be used reflexively, i.e. with a reflexive pronoun, in the same meaning as when it is used without a reflexive pronoun:

(3) Was erwarten Sie sich von dem Projekt?[15]
What expect you REFL from the project
'What are you expecting from the project?'

This usage is rare outside of Austria and South Tyrol. One is therefore tempted—based on hypothesis—to search for more instances of reflexive verbs in Austria (and South Tyrol) only. On the other hand, adopting a 'theory-agnostic' approach, like the one advocated in this paper, helps to ensure that no relevant data is overlooked. A case in point is the reflexive use of the verb *ausprobieren* 'to try', which ranked high in our metric when used with a reflexive pronoun.

14 http://www.krone.at/oesterreich/wahlbeteiligung-in-graz-sinkt-seit-1945-kontinuierlich-mangel-an-themen-story-341329 (10 February 2017).

15 http://www.nachrichten.at/oberoesterreich/wels/Gaesterekord-in-der-Vitalwelt-Bad-Schallerbach;art67,1059781 (8 February 2017).

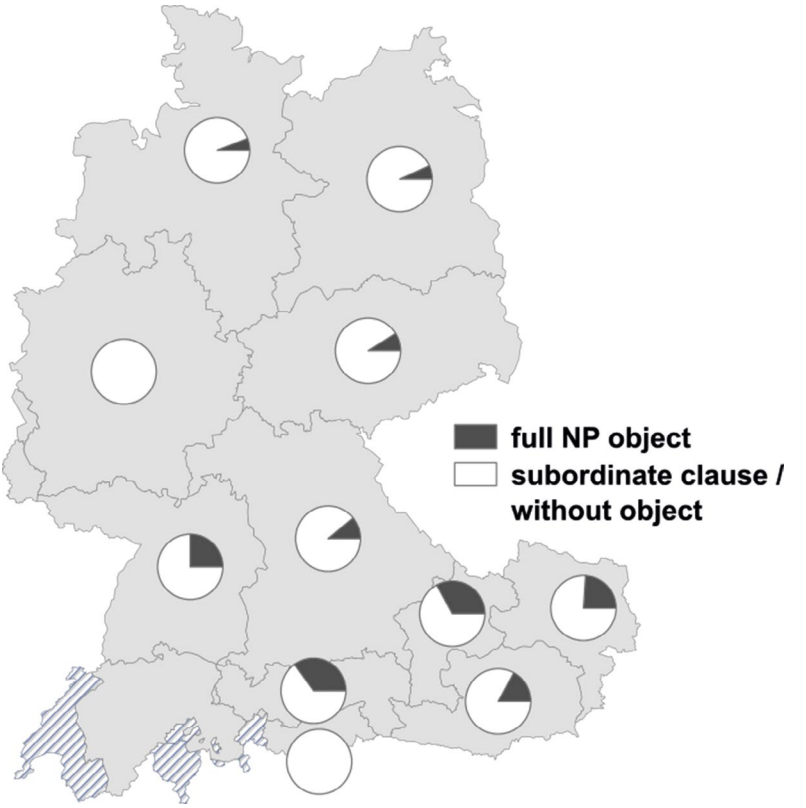


Figure 5: Distribution of *verlautbaren + full NP object* vs. *verlautbaren + with subordinate clause / without object.*

(4) In den Ferienkursen […] können sich Kinder ab zehn Jahren […]
in the holiday courses […] can REFL children from ten years […]
schauspielerisch ausprobieren.[16]
as-actors try-out
'In the holiday courses, children from the age of ten can dabble in acting.'

Sich ausprobieren (in/als) 'to try out something / to give something a try (in/as)' is used almost exclusively in Germany where it is most frequent in the subregions north-east (35 % of all hits in the corpus) and center-east (25 %). It is used less frequently in the other German subregions and in Belgium, and is hardly used in the other German-speaking countries/regions in Europe. To sum up: until very

16 http://www.tagesspiegel.de/berlin/gut-geruestet-fuer-die-freien-tage-unsere-freizeittipps-fuer-die-ferien/7717290.html (8 February 2017).

recently, the diatopically conditioned use of *sich ausprobieren* has not been documented.[17] We consider a semi-automatic approach promising for filling in gaps on the map of regional variation of German or, for that matter, of any language—gaps that tend to be overlooked in purely hypothesis-driven research settings.

From the point of view of variationist linguistics, the valency patterns presented in this section are clearly diatopically conditioned. At the same time, it is worth noting that these results cannot be interpreted in a strictly pluricentric model, i.e. a model where 'national varieties' are constitutive elements. National boundaries *are* an extralinguistic factor that can correlate with the diatopical distribution of variants in a standard language, but, at the same time, variation within or across national boundaries must be included systematically and without bias (cf. Niehaus 2015 as well as Elspaß and Dürscheid (2017) for discussion and references on *pluricentricity* vs. *pluriareality* in German).

4 Conclusion

This paper presented a semi-automatic method to identify regional grammatical variants. We discussed our pipeline approach that combines linguistic expertise and automatic ranking metrics and showed that it yields a fruitful combination in the sense that we discovered a reasonable number of (novel) variants while not having to go through too much noise (e.g. preprocessing errors) in the generated lists. We proposed an extended version of TF IDF which returned the most usable ranked lists containing variants with a substantial frequency in our corpus, while other metrics produced fewer variants or variants with less support in the corpus.

A theory-agnostic, (at least partially) data-driven approach like the one being put forward here is especially valuable in a field where ideologically colored discussions are common, even among linguists:

> "Offensichtlich wird die Diskussion um die Rolle der Arealität in der deutschen Standardsprache […] bisher eher politisch-ideologisch geführt" (Niehaus 2015: 138).
> 'It seems that the role of areality in the German standard language has been discussed in a rather political-ideological manner so far.'

17 The reflexive use of *ausprobieren* is mentioned neither in duden.de (last accessed: 8 February 2017)—as opposed to *sich versuchen in/als* '=' that is used in all German-speaking countries/regions, which *is* mentioned—nor in *Duden Zweifelsfälle* (2016), and *sich ausprobieren* was also not entered in the first edition of the *Variantenwörterbuch* (Ammon et al. 2004). However, it has been included in the second edition, where it is marked as "D", i.e. as a variant of Germany as a whole (Ammon/Bickel/Lenz et al. 2016: 69).

This paper contributes towards overcoming the lack of empiricism in research on variation within standard languages (cf. Niehaus 2015: 139).

References

Ammon, Ulrich et al. 2004. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol.* Berlin/New York: de Gruyter.

Ammon, Ulrich, Hans Bickel and Alexandra N. Lenz et al. 2016. *Variantenwörterbuch des Deutschen. Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen.* 2., völlig neu bearbeitete und erweiterte Auflage. Berlin/Boston: de Gruyter.

Bubenhofer, Noah. 2015. Muster aus korpuslinguistischer Sicht. In Christa Dürscheid and Jan Georg Schneider (eds.), *Handbuch Satz – Äußerung – Schema*, 485–502. Berlin/New York: de Gruyter.

Cao, Yan and Richard Xiao. 2013. A multi-dimensional contrastive study of English abstracts by native and non-native writers. *Corpora* 8.2: 209–234.

Clyne, Michael. 2004. Pluricentric Language. In Ulrich Ammon, Norbert Dittmar, Klaus J. Mattheier and Peter Trudgill (eds.), *Handbooks of linguistics and communication science.* Vol. 3: Sociolinguistics. 2., completely revised and extended edition, 296–300. Berlin/New York: de Gruyter.

duden.de A Website of Bibliographisches Institut GmbH. (8 February 2017).

[Dudengrammatik]. 2016. *Duden. Die Grammatik. Unentbehrlich für richtiges Deutsch.* 9., vollständig überarbeitete und aktualisierte Auflage. Berlin: Dudenverlag (= Duden 4).

[Duden Zweifelsfälle]. 2016. *Richtiges und gutes Deutsch.* 8., vollständig überarbeitete und erweiterte Auflage. Berlin: Dudenverlag (= Duden 9).

Dürscheid, Christa et al. In prep. *Variantengrammatik des Standarddeutschen. Ein Online-Nachschlagewerk.* Verfasst von einem Autorenteam unter der Leitung von Christa Dürscheid, Stephan Elspaß und Arne Ziegler.

Dürscheid, Christa and Stephan Elspaß. 2015. Variantengrammatik des Standarddeutschen. In Roland Kehrein, Alfred Lameli and Stefan Rabanus (eds.), *Regionale Variation des Deutschen – Projekte und Perspektiven*, 563–584. Berlin/Boston: de Gruyter.

Dürscheid, Christa and Patrizia Sutter. 2014. Grammatische Helvetismen im Wörterbuch. *Zeitschrift für Angewandte Linguistik* 60.1: 37–65.

Ebner, Jakob. 2008. Österreichisches Deutsch. Eine Einführung. Mannheim etc: Dudenverlag.

Elspaß, Stephan and Christa Dürscheid. 2017. Areale Variation in den Gebrauchsstandards des Deutschen. In Marek Konopka and Angelika Wöllstein (eds.), *Grammatische Variation – empirische Zugänge und theoretische Modellierung*, 85–104. Berlin/Boston: de Gruyter (= Jahrbuch des Instituts für Deutsche Sprache 2016).

Elspaß, Stephan and Konstantin Niehaus. 2014. The standardization of a modern pluriareal language. Concepts and corpus designs for German and beyond. *Orð og tunga* 16: 47–67.

Evert, Stefan 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, University of Stuttgart.

Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13.4: 403–437.

Laufer, Batia and Tina Waldman. 2011. Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning* 61.2: 647–672.

Mukherjee, Joybrato. 2009. The lexicogrammar of present-day Indian English. In Ute Römer and Rainer Schulze (eds.), *Exploring the Lexis-Grammar Interface*, 117–135. Amsterdam: John Benjamins.

Mukherjee, Joybrato and Sebastian Hoffmann. 2006. Describing verb-complementational profiles of New Englishes: A pilot study of Indian English. *English World-Wide* 27.2: 147–173.

Müller, Wolfgang. 2013. *Das Wörterbuch deutscher Präpositionen. Die Verwendung als Anschluss an Verben, Substantive, Adjektive und Adverbien*. Berlin/Boston: de Gruyter.

Niehaus, Konstantin. 2015. Areale Variation in der Syntax des Standarddeutschen. Ergebnisse zum Sprachgebrauch und zur Frage Plurizentrik vs. Pluriarealität. *Zeitschrift für Dialektologie und Linguistik* 82.2: 133–168.

Schmidlin, Regula. 2011. *Die Vielfalt des Deutschen. Standard und Variation. Gebrauch, Einschätzung und Kodifizierung einer plurizentrischen Sprache*. Berlin/New York: de Gruyter (= Studia Linguistica Germanica 106).

Schneider, Gerold and Lena Zipp. 2013. Discovering new verb-preposition combinations in New Englishes. *Studies in Variation, Contacts and Change in English* Vol. 13. http://www.helsinki.fi/varieng/journal/volumes/13/schneider_zipp/ (3 February 2017).

variantengrammatik.net Website of the research project *Regional Variation in the Grammar of Standard German* (14 February 2017).

Xiao, Richard. 2009. Multidimensional analysis and the study of world Englishes. *World Englishes* 28.4: 421–450.

Yoon, Hyung-Jo. 2016. Association strength of verb-noun combinations in experienced NS and less experienced NNS writing: Longitudinal and cross-sectional findings. *Journal of Second Language Writing* 34: 42–57.

Ziegler, Arne. 2010. ›Er erwartet sich nur das Beste …‹ Reflexivierungstendenz und Ausbau des Verbalparadigmas in der österreichischen Standardsprache. In Dagmar Bittner and Livio Gaeta (eds.), *Kodierungstechniken im Wandel. Das Zusammenspiel von Analytik und Synthese im Gegenwartsdeutschen*, 65–81. Berlin/New York: de Gruyter (= Linguistik – Impulse und Tendenzen 34).

Gosse Bouma

Corpus-Evidence for True Long-Distance Dependencies in Dutch

Abstract Long-distance dependencies have been studied extensively in syntactic theory. Yet, true long-distance dependencies, spanning more than a single predicate, appear to be rare in actual use. In this paper, we present the results of searching for such dependencies in a large, automatically annotated, treebank for Dutch, concentrating on phenomena that have recently been subject to debate, and where conflicting claims have been made regarding their productivity and existence.
Our results suggest that in Dutch, true long-distance dependencies are rare and have limited productivity. We also show that a popular strategy for avoiding such dependencies, resumptive prolepsis, is much more frequent and productive. Finally, we demonstrate that the annotation also facilitates searching for parasitic gaps, even though the construction itself is outside the scope of the computational grammar.

Keywords Long-distance dependencies, corpora, Dutch, resumptive prolepsis, parasitic gaps

1 Introduction

While syntactic theory has highlighted the possibility of potentially unbounded dependencies in WH-questions and relative clauses, in actual language use the dependencies introduced by a WH-question or relative clause are often very short and rarely span more than a single clause. To what extent genuine long-distance dependencies occur in natural language is therefore still an open question. Corpus-based research into this issue has been hindered by the fact that long-distance dependencies are difficult to find using search patterns consisting of lexical items and/or part-of-speech tags only. Syntactically annotated treebanks are more promising, as in theory they offer the kind of annotation required to identify long-distance dependencies. The Penn Treebank (Marcus et al. 1994) for instance, explicitly marks the relationship between WH-phrases

and relative pronouns and the 'extraction' site. However, carefully annotated and manually corrected treebanks are limited in size, while making claims about the possibility and productivity of certain long-distance dependencies requires corpora of considerable size. The alternative that we opt for in this paper is to work with automatically annotated data. The Alpino parser for Dutch (van Noord 2006) uses a linguistically motivated grammar and achieves high coverage and precision on most text genres.[1] The parser has been used to create the Lassy Large (van Noord et al. 2013), a large syntactically annotated corpus.

In this paper, we present the results of searching for four kinds of long-distance dependencies in an automatically annotated treebank for Dutch. We concentrate on phenomena that have recently been subject to debate, and where conflicting claims have been made regarding the question whether these constructions actually occur with some frequency in spontaneous language use. In particular, we will provide an answer to the following questions:

— To what extent do we find collocational effects in WH-questions and relative clauses involving a true long-distance dependency (Verhagen 2006)?
— To what extent do we find long-distance dependencies into infinitival clauses introduced by the optional complementizer *om*?
— What is the relationship between resumptive prolepsis (Hoeksema and Schippers 2012) and (the absence of) non-local dependencies?
— To what extent do we find parasitic gap constructions involving R-pronouns (Everaert et al. 2015) in actual text?

2 Background

One of the central topics in theoretical syntax is the proper analysis of non-local dependencies of the kind found in WH-questions and relative clauses. Rather different solutions have been proposed in various theoretical frameworks (among others in Transformational Grammar [Chomsky 1977], Categorial Grammar [Morrill 1995; Steedman 2000], GPSG [Gazdar et al. 1985], HPSG [Bouma et al. 2001], and LFG [Kaplan and Zaenen 1989]). One of the surprising facts is that there is still considerable disagreement about what the relevant data are and whether these are to be accounted for in syntax or by an appeal to general

1 In a recent comparison using the Universal Dependencies Lassy Small Corpus (http://universaldependencies.org/#nl_lassysmall), Alpino achieved labelled accuracy scores that were 4–7% higher than three state-of-the-art dependency parsers (including SyntaxNet) (Bouma and van Noord 2017).

cognitive constraints (Hofmeister and Sag 2010). Another observation that is somewhat at odds with the claims of most studies in theoretical syntax is that in actual usage, sentences involving a true long-distance dependency are rare, and often involve the same matrix verb and subject, suggesting that these are all variants of a small set of constructions (Verhagen 2006).

A corpus study can help to provide more insight in the frequency with which certain long-distance dependency constructions occur, and the amount of variation observed with each phenomenon. While WH-questions and especially relative clauses occur with some frequency in most corpora, cases that involve a true long-distance dependency (i.e. cases where the 'gap' is located in a subordinate clause) are not very frequent, and thus we will concentrate on material obtained from a large, but automatically parsed, corpus. This raises the question how accurate our results will be.

In computational linguistics, it has been observed that while statistical parsers now achieve very acceptable accuracies in general, this is not always the case when concentrating on more challenging aspects of syntax, such as properly accounting for non-local dependencies (Rimell et al. 2009; Candito and Seddah 2012). As we are using a corpus that was automatically annotated using the Alpino parser (van Noord 2006), this study can also give some insights into the accuracy of Alpino into analyzing non-local dependencies.

3 Non-local dependencies in the Lassy Corpus

The Lassy Large corpus (van Noord et al. 2013) is a corpus of contemporary Dutch that has been annotated with syntactic information. Annotation consists of lemmas, part-of-speech tags, constituent structure and dependency relations. It is composed of all material in the SONAR500 corpus (a mixed corpus of Dutch, containing texts from 18 different genres, i.e. adminstrative, autocues, magazines, legal, proceedings, web, etc., 41M sentences) (Oostdijk et al. 2013), Dutch Wikipedia (2011 dump, 9M sentences), EMEA (European Medicines Agency, 1M sentences), EUROPARL (proceedings of the European Parliament, 1M sentences), and various smaller sources. Syntactic annotation was done automatically using the Alpino parser (van Noord 2006). A small part of the corpus has been manually verified (Lassy Small, 65K sentences). Lassy Small and the Wikipedia-part of Lassy Large can be explored online.[2] In the examples below (Figure 1), we formulate queries using XPATH, as documented in Odijk (2015) and Augustinus et al. (2017).

2 http://zardoz.service.rug.nl:8067/

voor wie is de hulp bedoeld?
For who is the aid meant
'Who is the aid meant for?'

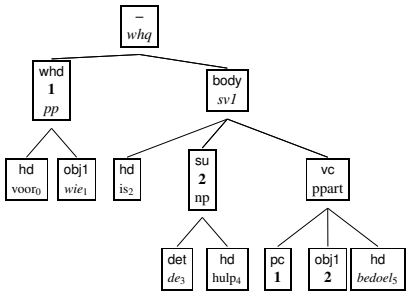


Figure 1: WH-question and corresponding syntactic dependency tree.

In this paper, we will be mostly concerned with syntactic constituency and dependency relations. As an example, consider the annotation of the WH-question sentence in Figure 1. The sentence initial WH-constituent *voor wie* is labeled with category PP. Internally, it consists of a head and a dependent labeled with the dependency relation `obj1` (used for objects of verbs and prepositions). The clause itself is a passive, headed by the auxiliary *is*, and containing two dependents: a subject and a verbal complement headed by a passive participle *(bedoeld)*. The passive participle phrase contains two empty nodes: a prepositional complement node co-indexed with the fronted PP and an object node co-indexed with the subject. The co-indexing between the initial PP and the prepositional complement of *bedoeld* expresses a non-local dependency. Following standard linguistic practice, we will sometimes refer to the latter type of node as a *'gap'*, even though the HPSG formalism on which the Alpino grammar is based does not actually employ gaps in its analysis of non-local dependencies.

Syntactically annotated corpora are useful for obtaining information about the distribution of such dependencies in actual usage. As a first example of how one can use a corpus to study non-local dependencies, we will look at the distribution of gaps in simple relative clauses. Simple finite clauses consist of a finite verb and one or more dependents that function as subject, direct object, indirect object, prepositional complement, etc. The dark bars in Figure 2 show that while all of these can be relativized, in 77% of the cases the gap is a subject. One might think that this is a consequence of the fact that subjects are simply more frequent than other dependents. The grey bars in Figure 2 show the distribution of all dependents in simple relatives (i.e. gapped or not). Only 37% of all dependents are subjects. This shows that in the vast majority of relative clauses, the gapped element is a subject, and that this preference is not (only) a consequence of the fact that in simple finite clauses, subjects are the most frequent dependents in general.

The statistics for gaps in simple relatives were obtained by running the following query on Lassy Small:

```
(1)  //node[ not(@word or @cat) and
               number(@index) = ../../node[@rel="rhd"]/number(@index)
           ]
```

This query searches for a node that has no `word`- or `cat`-attribute. This guarantees that the node does not correspond to a substring in the input sentence, i.e. it is a 'gap' (Figure 2).

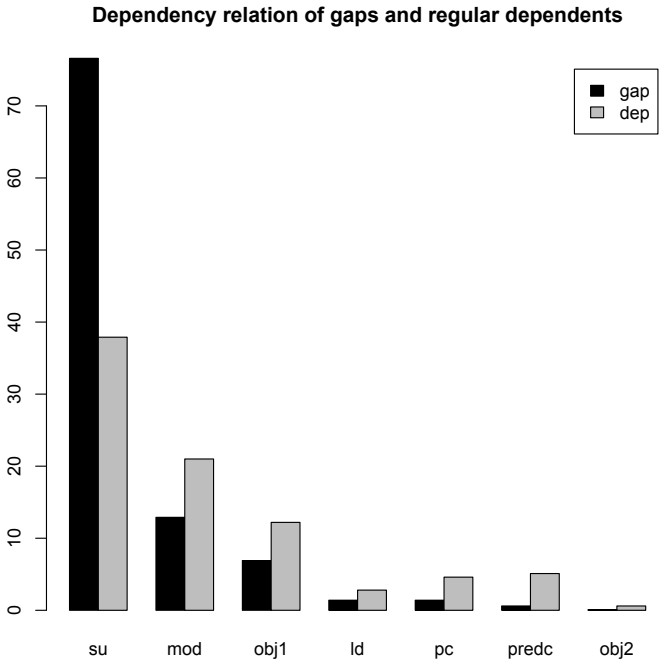


Figure 2: Distribution of dependency labels of gaps and regular dependents in simple relative clauses in Lassy Small.

Next, it requires that its `index` attribute has the same value as the node with dependency label `rhd` (this is the head of a relative clause), that occurs as a daughter (`'/node'`) of the grandmother (.../..) of the node itself. This ensures that we are only looking at 'local' instantiations of long-distance dependencies. It gives rise to over 8,000 hits.

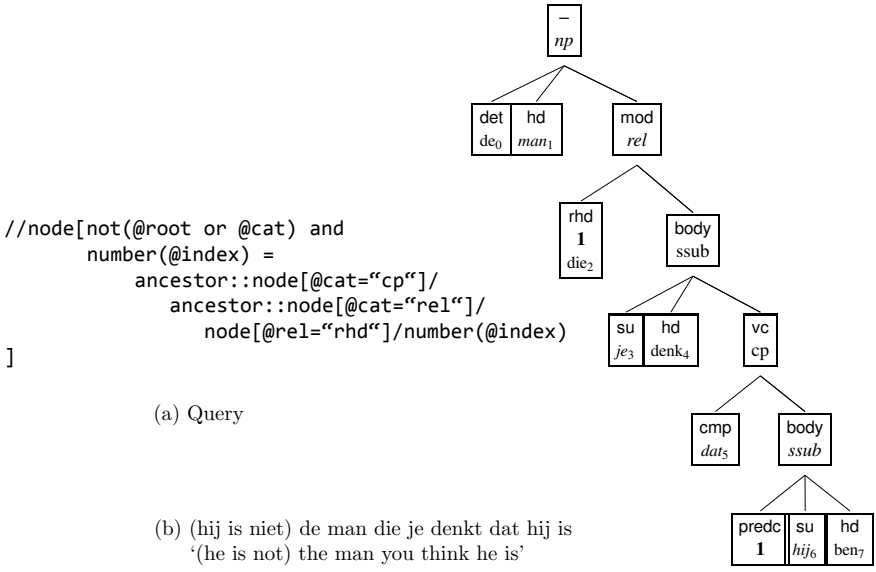
To obtain statistics for all dependents in the same set of relative clauses (the grey bars), we need to formulate a slightly more complex query:

```
(2)  //node[ not(@rel="hd") and
             ../node[ not(@word or @cat) and
                        number(@index) =
                           ../../node[@rel="rhd"]/number(@index)
                      ]
           ]
```

This query matches any non-head node that has a sister that meets the requirements of the previous query. Thus, we are looking at the same set of simple relative clauses as before, but now we can gather statistics for all non-head dependents (i.e. gapped or regular).

4 True long-distance dependencies

The dependency between a relative clause head and its corresponding gap is truly long-distance if the gap is located in a clause that is subordinate to the matrix verb of the relative clause or WH-question (Figure 3).[3]



(a) Query

(b) (hij is niet) de man die je denkt dat hij is
'(he is not) the man you think he is'

Figure 3: Long-distance dependencies in relative clauses.

3 Candito and Seddah (2012) use a slightly more liberal notion of true long-distance dependency that also includes 'gaps' in nominal and adjectival predicative phrases. Although such cases occur in Dutch, they are ignored in the present study.

There has been some discussion as to what extent such long-distance dependencies occur in (contemporary) Dutch, and whether they are limited to a small set of matrix verbs and subjects or not (Verhagen 2006; Hoeksema and Schippers 2012).

To find true LDDS in Lassy Large, we used the query in Figure 3a. It searches for a 'gap' dominated by a finite subordinate clause introduced by a complementizer[4] (i.e. its category is CP, for *complementizer phrase*), which in turn has to be dominated by a relative clause node (or WHQ node in the case of WH-questions). Furthermore, the index of the node has to be identical to the index of the head of the relative clause. An example of such a configuration is given in Figure 3b.

For the complete Lassy Large corpus, the query returned 270 hits for relatives, 73 of these were true LDDS (27 %). The query for WH-questions returned 2,601 hits, of which 344 cases were true LDDS (13 %). The distribution of matrix verbs in these examples is given in Table 1.

Table 1: Counts for matrix verbs in relative clauses and WH-questions with a true LDD.

Verb	N (rel)	N (wh)	Verb	N (rel)	N (wh)
denken ('to think')	52	252	hopen ('to hope')	3	1
willen ('to want')	7	49	weten ('to know')	2	0
zeggen ('to say')	4	5	vermoeden ('to suspect')	2	0
vinden ('to find')	3	33	zien ('to see')	0	1
			wensen ('to wish')	0	1
			verwachten ('to expect')	0	2

The dominance of *denken* is striking, and confirms to some extent the observations in Verhagen (2006).

It should also be noted however, that the corpus contains a fair amount of user generated content from social media. In this text genre, the relative clause *die je/hij/ik/ze denk(t)(en) dat je/hij/ik/ze is/ben (that you think I am* and pronominal variants) is a frequently occurring phrase.

Recently, there has been quite a bit of discussion about the possibility of *weten* as matrix verb in long-distance dependency constructions (Coppen 2013).[5] It has been claimed that only non-factive verbs can be matrix verbs in long-distance dependencies of this kind (Ross 1967). Coppen points out that similar examples involving *weten* can be found relatively easily in literature from the 17th and 18th century, and also suggests that *weten* might not be strictly factive

4 Note that in Dutch, the presence of a complementizer is obligatory in this construction.

5 The discussion in the media was triggered by the phrase *de dag die je wist dat zou komen (the day that you knew that would come)* from a song composed on the occasion of the coronation of King Willem Alexander (2013).

in all contexts. Our results show that even in modern Dutch, the use of *weten* in true LDDS is not completely excluded. These are the two examples with factive matrix verb *weten*:

(3) a. ik ben nog steeds niet de volwassene die ik wist dat ik kon zijn
I am still still not the adult that I knew I could be
'I am still not the grown-up that I knew I could be'

b. ik pak alleen mensen die ik weet da eerlijke kans maken
I grab only people that I know that honest chance make
'I only attack people that I how have an honest chance'

Verhagen (2006) finds that in his corpus (Eindhoven corpus and articles from 'de Volkskrant'), the subject in WH-questions involving a long-distance dependency is almost always a second person pronoun. The distribution in the examples found in true LDDS in the Lassy corpus (Table 2) confirms that this is indeed predominantly the case for WH-questions. For relative clauses, however, a more diverse picture emerges. There is a strong preference for pronominal subjects, but first, second, and third person pronouns are all of approximately the same frequency.

Table 2: Distribution of subjects in matrix clauses in true LDDS.

	Relatives	WH-questions
first person	25	9
second person	23	313
third person pronouns	25	11
full NPs	3	13
other	2	

The Alpino grammar specifies lexically which verbs that take a clausal complement can occur as matrix verbs in long-distance dependency constructions (these verbs are sometimes called *'bridge verbs'*). This list is slightly larger than the verbs mentioned in Table 1, and also contains *bedoelen* ('to mean'), *beloven* ('to promise'), and *beweren* ('to claim'). Even if longdistance dependencies are rare, the size of the Lassy Large corpus would lead one to expect that at least for all of these verbs, some examples can be found. Of course, we should keep in mind that the Lassy Large corpus was automatically analyzed and thus some relevant cases may have been missed. For instance, manual inspection of all relatives with matrix verb *beweren* and containing a subordinate clause in the Wikipedia section of Lassy Large did reveal one case involving a long-distance dependency:

(4) de naam waaronder men beweerde dat Menelaos een tempel
the name under-which one claimed that Menelaos a temple
voor Aphrodite had opgericht
for Aphrodite had founded
'the name under which one claims that Menelaos had founded a temple for Aphrodite'

Of all the question sentences with matrix verb *beweer* in Lassy Large (116 cases), not a single one contained a true LDD. Also, manual inspection of all WH-questions with *bedoelen* and *beloven* as matrix verb did not return a single case with a true LDD. It is thus not impossible that examples of true LDDS involving other 'bridge' verbs are present in the corpus, but at the same time these results suggest that they will not be very frequent.

True LDDS are extremely rare in the Lassy Large corpus. For a similar construction in English, relatives involving subject extraction from an embedded clause, Rimell et al. (2009) report that it occurs in 0.4 % of the sentences in their corpora (Wall Street Journal and Brown). The Lassy Large corpus contains more than 50M sentences, and thus even if the recall of the Alpino parser is low on this phenomenon, it seems unlikely that more than several thousand (i.e. 0.002–0.01%) of the sentences in Lassy Large contain a true LDD.

5 Long distance dependencies with non-finite clauses

It is not exactly clear what should be counted as a long-distance dependency. Usually, cases involving an auxiliary or modal as in (5) are not seen as long-distance, even though one might claim that these involve a matrix clause (the auxiliary or modal and the subject) and an embedded non-finite VP.

(5) de kiesdrempel die de partij zelf had ingevoerd
the election-threshold that the party itself had introduced
'the election threshold that the party had introduced itself'

However, there are also verbs that select a *to*-infinitival complement, where the matrix verb cannot be seen as a modal or auxiliary (Cremers 1983). In those cases where the to-infinitival complement is in 'extraposed' position, it can be optionally introduced by the complementizer *om:*

(6) De stichting is verplicht (om) haar winst aan sociale projecten
The foundation is obliged (CMP) her profit to welfare projects
uit te keren
out to turn
'The foundation is obliged to give her profit to welfare projects'

It seems reasonable to categorize relative clauses that involve a dependency with a gap inside a *to*-infinitive of this kind as true LDDs as well. An interesting question in this case is the role of the optional complementizer. The presence or absence of *om* is influenced by various factors involving sentence complexity, such as distance between the matrix verb and complement, frequency of the matrix verb, and frequency with which the matrix verb occurs with a VP-complement (Bouma 2017). Whether the presence of a long-distance dependency also influences the likelihood of the complementizer *om* is unclear. For instance, Bennis (2000) presents example (7-a), where *om* is marked as optionally possible. Broekhuis et al. (1995) present example (7-b), but add in the discussion that *'it must be mentioned that the complementizer is preferably dropped'*.

(7) a. Waar is Jan bang (om) over te praten
Where is John afraid (CMP) over to talk
'What is John afraid of to talk about'
b. Wat heeft Jan geprobeerd om te lezen
What has John tried CMP to read
'What has John tried to read'

We tried to find cases like this in the corpus. The search for cases that are introduced by *om* is relatively straightforward, and requires only a minor variation of the query given above for finite complements (i.e. instead of a node with category CP we now search for the same configuration with a node of category OTI (for *om-te*-infinitive)):

(8) een boek dat je intellect simpelweg weigert om serieus te nemen
a book that your intellect simply refuses CMP seriously to take
'a book that your intellect simply refuses to take seriously'

When searching for cases where the complementizer is absent, we added an additional constraint to the query that requires that the *te*-infinitive contains at least one dependent that follows the matrix verb but precedes the verb heading the infinitival clause, as in (9-a). This ensures that the infinitive is indeed an '*extraposed*' complement, and has not been integrated into the matrix clause as a

result of a process that is known as *'verb raising'*, as in (9-b). In the latter case, it is unclear whether there is indeed a long-distance dependency.

(9) a. organisaties die ik vergeten ben een adreswijziging te sturen
organisations that I forgot am an address-change to send
'organisations to which I forgot to send a change of address'
b. organisaties die ik een adreswijziging ben vergeten te sturen

The results for searching for true LDDS in infinitival complements are given in Table 3. There is quite a bit of variation in matrix verbs in both cases (16 different types for *om-te*-infinitives, and 22 different types for *te*-infinitives). The only verb that occurs with a high frequency (21 hits) is *achten* ('to suppose') in the *te*-infinitive case, as in (10). This is unexpected, as *achten* is not a very frequent verb in general.

Table 3: Counts for true LDDS involving infinitival complements

	hits	valid	verb types
om-te-infinitives	81	28	16
te-infinitives	275	75	22

(10) conversaties die ze geacht worden niet te horen
conversations that they supposed are not to hear
'converstations that they were not supposed to hear'

Our results confirm that true LDDS are possible with both *om-te*-infinitives and *te*-infinitives, and that this is possible for a wide range of matrix verbs. The results do not give a clear answer to the question whether true LDDS are less likely if *om* is present, as the two data-sets are not very comparable (i.e. we added an additional constraint to the query for *te*-infinitives).

Manual checking was necessary to obtain the results in this section and the preceding section. As a result, we can observe that the precision of the Alpino parser on true LDDS in relative clauses in Lassy Large is 35% (73/209), 13% (344/2601) for true LDDS in questions, 35% for *om-te*-infinitives (28/81) and 27% for *te*-infinitives in extraposed position. This may not seem very high, but, with the exception of WH-questions, it is in fact comparable to the performance of the best performing system in Rimell et al. (2009) on subject extraction from an embedded clause. It should also be noted that these make up a tiny portion of the corpus as a whole, and thus, the effect on parser accuracy in general is negligible.

6 Resumptive prolepsis

Hoeksema and Schippers (2012) present results from a diachronic corpus study suggesting that true LDDs are in decline in Dutch, and that, especially in relative clauses, they are being replaced by a construction referred to as *'resumptive prolepsis'* by Salzmann (2006) and which involves a relative clause headed by *waarvan* ('of which') or *van wie* ('of whom') and a 'resumptive' pronoun in an embedded clause:

(11) a. 45 mogelijke van Goghs **waarvan** onduidelijk is of **ze**
45 potential van Gogh's of-which unclear is whether they
echt of vals zijn
true or fake are
'45 potential van Gogh's of which it is unclear whether they are true or false'

b. iemand van wie ze denkt dat hij haar man is
somebody of-which she thinks that he her husband is
'somebody that she thinks is her husband'

The Alpino parser does analyse these as relative clauses where the relative head is co-indexed with a gap in the matrix clause that is labeled as a modifier. It does not establish a relation between the pronoun in the subordinate clause and the relative clause head. To find instances of this construction involving the adverbial PP *waarvan*, we used the following query:

```
(12) node[ @cat="rel" and node[@lemma="waarvan"]]/
        node[ .//node[@rel="mod" and @index]]//
            node[ @cat="cp" and (@rel="su" or @rel="vc")]//
                node[ @pt="vnw" and (@rel="su" or @rel="obj1") and
                      (@vwtype="pers" or @vwtype="aanw") ]
```

This query searches for relative clauses headed by *waarvan*, dominating a node that has a descendant that is an indexed modifier (the gap) and which has a descendant that is a finite subordinate clause with dependency label su or vc. The latter constraint ensures that the CP is indeed a complement, and not a modifier. Finally, the subordinate clause has to contain a personal or deictic pronoun with dependency label su (for subject) or obj1 (for direct object, of a verb or preposition). The query for *van wie*-cases is similar except for the definition of the relative clause head.

This query, while only approximating the requirements of the resumptive prolepsis construction, returns more than 9,500 hits and turns out to be quite

Table 4: Distribution of matrix verbs and pronouns in *waarvan/van wie*$_i$*... pronoun*$_i$ constructions

matrix verb	hits	%	pronoun	hits	%
weten ('to know')	1489	15.6	ze	3537	37.1
denken ('to think')	1324	13.9	het	2340	24.6
bekend zijn ('be known')	851	8.9	hij	1282	13.4
zeggen ('to say')	709	7.1	die	752	7.9
vermoeden ('to suspect')	498	5.2	zij	729	7.9
hopen ('to hope')	396	4.2	deze	581	6.1
verwachten ('to expect')	392	4.2	er	280	3.0
vinden ('to .nd')	376	3.8	hem	134	1.4
veronderstellen ('to suppose')	254	2.6	dat	117	1.2
beweren ('to claim')	249	2.6	dit	91	1.0
other	3413	34.3	*other*	1.0	

accurate. In a random sample of 100 sentences, we found only 4 false hits, suggesting a precision of 96%. Most cases (8,031) are with *waarvan* as relative clause head, 1,490 have *van wie* as relative clause head. The complement clause is usually a regular verbal complement, but sometimes (1,488 cases) functions as subject. The complementizer is almost always *dat* (9,062 cases), but examples with complementizer *of* and *alsof* occur as well (459 cases).

The distribution of matrix verbs and matching resumptive pronouns is given in table 4. The two most frequent verbs are *denken*, which is most frequent for true, and *weten*, for which it is usually claimed that it cannot occur in long-distance dependencies. The data confirms the observation in Hoeksema and Schippers (2012) that this construction is not subject to island constraints: there is a wide variety in matrix verbs, most of which are not known to be 'bridge verbs', in 459 cases the resumptive pronoun is in a complement clause headed by *(als)of,* and in 1,488 the resumptive pronoun is in a subject clause. The latter are mostly cases involving the copula *zijn:*

(13) Soorten **waarvan** het onduidelijk is of **ze** in Nederland voorkomen
species of-which it unclear is whether they in the Netherlands occur
'species of which it is unclear whether they occur in the Netherlands'

7 R-Pronominal Parasitic gaps

In the previous sections we have been concerned with searching for true LDDs in an annotated corpus, and searching for a popular strategy for avoiding such dependencies. In this section we add some observations on a closely related construction that seems to be extremely scarce in actual data as well.

In Dutch, non-local dependencies between a fronted WH-element and a position governed by a preposition are in general not allowed. So-called *'R-pronouns'* (following the discussion in van Riemsdijk [1978]) are an exception to this rule. They can be used both to form WHquestions, as in (14-b), as well as discontinuous constituents where the r-pronoun precedes but is non-adjacent to its governing preposition (14-d).

(14) a. *Wat ben je voor verzekerd?
What are you for insured
b. Waar ben je voor verzekerd?
What[+R] are you for insured
'What are you insured for¿
c. *Je bent het niet voor verzekerd
You are it not for insured
d. Je bent er niet voor verzekerd
You are it[+R] not for insured
'You are not insured for it'

A recent paper (Everaert et al. 2015) arguing for structure being more prominent than word order in syntax uses this construction to produce Dutch example sentences like (15-b).

(15) a. Ik ben speciaal voor het klimaat naar de Provence toe gereden
I am especially for the climate to the Provence driven
'I drove to Provence especially for the climate'
b. Ik ben **er** speciaal **voor naar toe** vertrokken
I am it especially for to to driven
'I drove there especially for it'

Compared to (15-a), which does contain two full PPs, the R-pronoun *er* in (15-b) seems to be dependent on a gap in two PPs. Everaert et al. (2015) draw a parallel between cases such as this and parasitic gap constructions (Engdahl 1983). The examples were discussed in a blog[6] that sparked a lively discussion, including a response by one of the authors of the original paper.[7]

6 http://nederl.blogspot.nl/2015/11/ik-ben-er-speciaal-voor-naartoe-gereden.html
7 http://nederl.blogspot.nl/2015/11/recursie-en-evolutie-van-taal.html

While this construction does not involve a true long-distance dependency, we include it in our discussion as it does involve a rare construction involving non-local dependencies.

Huijbrechts (p.c., Huijbrechts [2016]) presents additional examples such as (16).

(16) a. **Waar** rekent hij **op** om **naar toe** te gaan?
Where counts he on PRT to to to go
'Where does he count on to go to?'
b. **Waar** ga je **van** uit dat zij **op** zal letten?
Where go you from out that she on will note
'What do you suppose she will pay attention to?'

These constructions are a slight variation of the R-pronominal parasitic gap constructions in (15-b), in that they involve a gap in a PP in a complement clause, and a suppressed R-pronoun in the main clause. Note that normally, PPs containing a complement clause are obligatorily introduced by the expletive R-pronoun er:

(17) a. Hij rekent er op om naar Amsterdam toe te gaan
He counts there on CMP to Amsterdam to to go
'He counts on going to Amsterdam'
b. Je gaat er van uit dat zij op schrijffouten zal letten?
You go there of out that she on spelling-errors will notice
'You are counting on her to pay attention to spelling errors'

One of the questions is to what extent such phenomena occur in spontaneous data. If not, or scarcely, they constitute evidence for a 'Poverty of the Stimulus' argument: apparently, language users are able to produce and understand parasitic gap constructions without necessarily having been exposed to such sentences in the past.

One problem with this argument is that it is very hard to check for the occurrence of configurations such as (15-b) and (16) in corpora. The Alpino parser, while based on a linguistically sophisticated hand-written grammar, does not cover parasitic gap constructions. As a consequence, these will not be analyzed as such in corpora that are analyzed automatically by Alpino. Given a sufficiently large corpus, one might search for sentences containing the trigram *voor naar toe* and check these manually. The NL-COW corpus (text from Dutch language websites, 259M sentences)[8] contains 19 occurrences of the string *voor naartoe*,[9] of which at least a few cases are similar to the example presented by Everaert et al. (2015):

8 http://corporafromtheweb.org
9 We opted for searching for the more common spelling *naartoe over naar toe.*

(18) a. … ik zou **er** niet speciaal voor naartoe gaan
… I would there not especially for towards go
… 'I would not especially go there for it'
b. **Er** speciaal **voor naartoe** rijden hoefde niet
There especially for towards drive needed not
'It was not necessary to drive there for it especially'

However, this kind of search is very limited, as (1) it presupposes that the two prepositions are adjacent, which need not be the case in parasitic gap constructions in general, and (2) it fails to check for cases involving other prepositions.

Another possibility is spotting such constructions 'in the wild'. For instance, after becoming aware of examples such as (16), we noticed the following quote:[10]

(19) Daar heb je dan geen tijd voor om naar te kijken
there have you than no time for cmp to to watch
At that moment, you do not have time to look at that

This suggests that maybe constructions like these have simply gone unnoticed by linguists.

A more effective strategy involves searching for potential parasitic gaps in Lassy Large. As Alpino does not take parasitic gaps into account, we will have to formulate a query that only approximates the relevant syntactic configuration, and check results manually. We used the following query:

(20)
```
//node[node[@rel="rhd" and @lemma="waar"] and
    descendant::node[node[@cat="pp"]/node[@index and not(@pos or @cat)]
                     descendant::node[@rel="vc" and
                                      (@cat="ti" or @cat="oti" or @cat="cp")
                                          ]
                    ]
      ]
```

Here, we search for sentences containing a relative clause headed by *waar*, and containing a PP containing a gap, and a complement clause. Such sentences might, but are not guaranteed to, contain the relevant structure.

The query gives rise to 564 hits on Lassy Large, of which 16 cases appear to be instances of the phenomenon we are interested in. Two examples are given below:

10 Interview with cyclist Matteo Trentin (translated into Dutch) by Nando Broers in *De Muur*, 2016/2.

(21) a. Het soort waar iedere vrouw van zou moeten dromen
The kind of-which every woman of should must dream
om te trouwen
COMP to marry
'the kind which every woman should dream of to marry with'

b. Dit zijn de genen waar men voor heeft gekozen om onderzoek
These are the genes which one for has chosen COMP research
naar te doen
into to do
'These are the genes for which one has chosen to do research on'

The results of the query are very noisy. Although it may be possible to modify the query to achieve slightly better precision, we do believe that these constructions are very hard to detect in the output of the current Alpino grammar. In terms of frequency, examples like these do seem almost as frequent as long-distance dependencies in relative clauses containing a gap in a tensed subordinate clause or in a complement clause introduced by *om*.

8 Conclusions

In this paper, we have searched for true long-distance dependencies in an annotated corpus. True LDDs in relatives and WH-questions containing a subordinate clause (either tensed or introduced by the complementizer *om* or containing an 'extraposed' infinitival complement) are all covered by the Alpino parser, and thus can be searched for directly. Manual inspection of the results was necessary as the precision of the parser on these constructions is not very high. The results show that true LDDs are quite infrequent in the corpus but do seem to provide support for claims that there are collocational effects in this construction.

Two related constructions, resumptive prolepsis and R-pronominal parasitic gaps, are outside the scope of the grammar. For the resumptive pronoun construction, an approximate query turned out to be quite accurate, and gave rise to a high number of results. The distribution of matrix verbs in this construction supports the findings of Hoeksema and Schippers (2012). For R-pronominal parasitic gaps, it is much harder to come up with a good approximate query. However, after manual filtering we did find a number of positive examples. In this case, the main advantage of using a syntactically annotated corpus is that it makes it possible to search somewhat efficiently for this phenomenon in the first place.

The Lassy Large corpus seems sufficiently large and heterogeneous to support research on long-distance dependencies, and the automatic syntactic annotation,

while far from perfect, does help to zoom in on the interesting cases quickly. Several questions remain for further research, such as estimating the recall of the automatic parser, and collecting statistics for other longdistance dependency constructions, such as comparatives.

References

Augustinus, Liesbeth, Vandeghinste, Vincent, Schuurman, Ineke and van Eynde, Frank. 2017. GrETEL: A tool for example-based treebank mining. In Jan Odijk and Arjan van Hessen (eds.), *Clarin in the Low Countries*, 269–280. London: Ubiquity Press.

Bennis, Hans. 2000. Adjectives and argument structure. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, 27–68.

Bouma, Gosse. 2017. Om-omission. In Martijn Wieling, Martin Kroon, Gosse Bouma and Gertjan van Noord (eds.), *From Semantics to dialectometry: Festschrift for John Nerbonne*, 65–74. London: College Publications.

Bouma, Gosse, Malouf, Rob and Sag, Ivan. 2001. Satisfying Constraints on Adjunction and Extraction. *Natural Language and Linguistic Theory* 19, 1–65.

Bouma, Gosse and van Noord, Gertjan. 2017. Increasing return on annotation investment: the automatic construction of a Universal Dependency treebank for Dutch. In Joachim Nivre and Marie-Catherine de Marneffe (eds.), *NoDaLiDa workshop on Universal Dependencies*, Gothenburg.

Broekhuis, Hans, Den Besten, Hans, Hoekstra, Kees and Rutten, Jean. 1995. Infinitival complementation in Dutch: On remnant extraposition. *The Linguistic Review* 12(93-122).

Candito, Marie and Seddah, Djam'e. 2012. Effectively long-distance dependencies in French: annotation and parsing evaluation. In Iris Hendrickx, Sandra Kübler and Kiril Simov (eds.), *TLT 11 – The 11th International Workshop on Treebanks and Linguistic Theories*.

Chomsky, Noam. 1977. On wh-movement. In Akmajian Adrian Culicover Peter, Wasow Thomas (ed.), *Formal Syntax*. New York: Academic Press.

Coppen, Peter-Arno. 2013. De zin die wij merken dat ook voor linguïstische problemen zorgt. *Nederlandse Taalkunde* 18(2), 193–203.

Cremers, Crit. 1983. On two types of infinitival complementation. In Frank Heny (ed.), *Linguistic categories: auxiliaries and related puzzles*. 169–221, Dordrecht: Springer.

Engdahl, Elisabet. 1983. Parasitic gaps. *Linguistics and philosophy* 6(1), 5–34.

Everaert, Martin, Huybregts, Marinus, Chomsky, Noam, Berwick, Robert and Bolhuis, Johan. 2015. Structures, not Strings: Linguistics as part of the Cognitive Sciences. *Trends in Cognitive Sciences* 19, 729–743.

Gazdar, Gerald, Klein, Ewan, Pullum, Geoffrey and Sag, Ivan. 1985. *Generalized Phrase Structure Grammar*. Oxford: Blackwell.

Hoeksema, Jack and Schippers, Ankelien. 2012. Diachronic changes in long-distance dependencies. *Historical Linguistics 2009: Selected Papers from the 19th International Conference on Historical Linguistics, Nijmegen, 10–14 August 2009,* 155–170.

Hofmeister, Philip and Sag, Ivan A. 2010. Cognitive constraints and island effects. *Language* 86 (2), 366–415.

Huijbrechts, Riny. 2016. Binding Unleashed. Ms. Utrecht University.

Kaplan, Ronald M. and Zaenen, Annie. 1989. Long-distance Dependencies, Constituent Structure and Functional Uncertainty. In Mark R. Baltin and Anthony S. Kroch (eds.), *Alternative Conceptions of Phrase Structure*, University of Chicago Press.

Marcus, Mitchell P., Santorini, Beatrice and Marcinkiewicz, Mary Ann. 1994. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19(2), 313–330.

Morrill, Glyn. 1995. Discontinuity in categorial grammar. *Linguistics and Philosophy* 18(2), 175–219.

Odijk, Jan. 2015. Linguistic Research with PaQu. *Computational Linguistics in The Netherlands journal* 5, 3–14.

Oostdijk, Nelleke, Reynaert, Martin, Hoste, Véronique and Schuurman, Ineke. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In Peter Spyns and Jan Odijk (eds.), *Essential speech and language technology for Dutch*, 219– 247. Heidelberg: Springer.

Rimell, Laura, Clark, Stephen and Steedman, Mark. 2009. Unbounded dependency recovery for parser evaluation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Volume 2, 813–821, Association for Computational Linguistics.

Ross, J.R. 1967. *Constraints on variables in syntax*. Ph. D.thesis, Massachusetts Institute for Technology.

Salzmann, Martin. 2006. *Resumptive Prolepsis: A Study in Indirect A' Dependencies*. Ph. D.thesis, Leiden University, Leiden.

Steedman, Mark. 2000. Information structure and the syntax-phonology interface. *Linguistic inquiry* 31(4), 649–689.

van Noord, Gertjan. 2006. At Last Parsing Is Now Operational. In Piet Mertens, Cedrick Fairon, Anne Dister and Patrick Watrin (eds.), *TALNo6. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, 20–42.

van Noord, Gertjan, Bouma, Gosse, van Eynde, Frank, de Kok, Daniel, van der Linde, Jelmer, Schuurman, Ineke, Sang, Erik Tjong Kim and Vandeghinste, Vincent. 2013. Large Scale Syntactic Annotation of Written Dutch: Lassy. In

Peter Spyns and Jan Odijk (eds.), *Essential Speech and Language Technology for Dutch: the STEVIN Programme*, 147–164, Springer.

van Riemsdijk, Henk. 1978. *A Case Study in Syntactic Markedness: The binding nature of prepositional phrases*. Foris Publications, Dordrecht.

Verhagen, Arie. 2006. On subjectivity and 'long distance Wh-movement'. In Angeliki Athanasiadou, Costas Canakis and Bert Cornillie (eds.), *Subjectification: Various Paths to Subjectivity*, 323–346, Berlin: Mouton de Gruyter.

Yela Schauwecker, Achim Stein

Automatic Morphosyntactic and Dependency Annotation of the Anglo-Norman Text Database

Abstract Non-standardized languages are an immense challenge for automatic annotation. This paper discusses the case of Anglo-Norman (AN), which is the variety of Old French (OF) spoken and written in medieval England for over 300 years, until well after 1400. In addition to presenting the irregularities in, for example spelling, inflection and word-order that are also characteristic of OF, AN developed particular spelling variants, shows even less consistent case-marking and considerable diachronic variation between the earliest (c1112) and the latest (c1440) texts in the Anglo-Norman text database (Rothwell and Trotter 2005; henceforth "ANdb").

We present the first attempt to provide an automatic grammatical analysis of the ANdb. We applied machine-learning techniques combined with lexicon-driven tools that were trained on OF resources. This paper is organized according to the individual steps in the annotation process: section 1 gives a succinct overview of the historical context and some relevant linguistic peculiarities of AN. Section 2 deals with the automated graphical "normalisation" of the texts. We generated regularized spellings that temporarily substituted the graphical forms during the annotation process to improve the accuracy of lemmatisation, part-of-speech tagging, and dependency parsing. Section 3 describes how a dependency parser developed for Old French was applied to the normalised version of the AN data, and discusses the usefulness of the parsed output for historical syntactic research.

Keywords Dependency parsing, part of speech tagging, automatic spelling normalisation, Anglo-Norman, Old French historical corpora

1 Anglo-Norman

1.1 Timeline of French in England

When William the Conqueror arrived at Pevensey in 1066, he brought with him the variety of Old French (OF) that was spoken in Normandy. At the beginning, Norman OF was the dominant code in England, which influenced the less prestigious Middle English. But, a few generations later, French speakers were almost always mother-tongue speakers of English, so that Insular French was maintained by largely fluent bilinguals (Ingham 2012). In contrast to earlier assumptions, Ingham found evidence that Anglo-Norman (AN)[1] showed no signs of decline until the fifteenth century (see also Hunt 2004). Since evidence for the systematic teaching of French emerges only just before that point, the acquisition of French by anglophone speakers until then must have taken place via natural interaction with French speakers.

1.2 Some features of Insular French

Knowing the syntactical features of AN, and in particular those that set AN apart from (continental) OF, is crucial to understanding the additional difficulties automatic annotation has to cope with in the case of insular texts. However, their detailed description is beyond the scope, and the topic, of this paper. Therefore, we just give some examples for the sake of illustration.

Being originally a variety of OF, AN shares most of the characteristic features of this language. Among these, the absence of a standardized spelling, inconsistent word-order, the licensing of null-subjects (see Marchello-Nizia 2009, among others), all represent major difficulties in automatic linguistic annotation. However, our tools are trained on OF resources (see section 3), and therefore, it is on OF that they achieve best results.

When it comes to Anglo-Norman, the situation gets more difficult. Even bare numerical comparison can reveal the high level of syntactic complexity in AN compared to OF: texts in the *Syntactic Reference Corpus of Medieval French (SRCMF)*[2] contain 24,171 "sentences" within 266,870 tokens, thus equalling an average of 11.04 words per sentence. Compared to that, texts in the ANdb contain

1 A number of researchers prefer the term "Anglo-French". We agree, but because of the more technical scope of this contribution, and in order to avoid confusion, we will use "Anglo-Norman" throughout this text.

2 Calculations based on the version 0.91, March 8, 2016.

3,111,982 in 148,353 "sentences"[3], which equals an average of 21 words per "sentence". In addition to that, AN was spoken and written for about 400 years, and therefore shows much diachronic variation in itself between the first (c1112) and the latest texts in the ANdb (c1440).

Like OF, AN showed considerable graphical irregularity from the start. But in the case of AN, these irregularities increased considerably, as AN phonology underwent some profound changes by the later thirteenth century. Phonological contrasts that had been kept up in earlier times ceased to be respected by later generations of speakers (Ingham 2012: 160). This is, of course, at least partly reflected in the orthography. In addition, AN exhibits a number of atypical traits by which it is set apart from continental OF (Ingham 2010), many of which are highly relevant to syntactic annotation. For example, the contrast between strong and weak forms of pronouns ceases to be respected in many cases (Grant 1978: 36–7; Johnston 1961:xix; Ingham 2010), and direct and indirect object case-marking is confused in later texts (Grant 1978: 36, Johnston 1961: xix; Ingham 2010). To summarise, as these examples illustrate, AN diverges from OF in syntax as well as in phonology and orthography. As a consequence, there is a clear difference between the texts our tools were trained on and the AN sources they are applied to. The following sections illustrate the approaches we adopted in order to bridge this gap and the results we achieved.

1.3 Pre-processing of the Anglo-Norman text database (ANdb)

The Anglo-Norman text database was compiled in order to support the *Anglo-Norman Dictionary* project (AND, Rothwell and Trotter 2005). It is freely accessible on the internet via the *Anglo-Norman On-Line hub* (ANHub[4]). It contains 78 texts, from c1112 to c1440.[5] At this point it must be noted that providing a fully annotated version of the ANdb is clearly beyond our possibilities, as is often the case with low-resourced but richly documented languages. However, in the case of AN, additional difficulties have to be dealt with. As we said above, the enormous syntactic complexity of especially the later AN documents – a considerable amount of "sentences" contains 200 and more tokens – would make full annotation extremely time-consuming and error-prone. In addition to that,

3 As to the notion of „sentence" in the SRCMF cf. infra, section 1.3.

4 http://www.anglo-norman.net.

5 The data used for the annotation presented here were kindly provided by Geert de Wilde within a research collaboration between the *Anglo-Norman Dictionary* project (AND) and the project *Borrowing of Argument Structure in Contact Situations* (BASICS), funded by the Deutsche Forschungsgemeinschaft 2015–2018.

texts often contain English, French and Latin words all within the same sentence. Thus, annotating them represents a major challenge even for well-trained human annotators, let alone elaborating a verified "gold-standard" version of the corpus. Moreover, as to the texts themselves, it has to be taken into account that the ANdb is heterogeneous in many respects: it contains prose as well as verse-texts, dating from very different periods and reflecting very different states of the language. They deal with an immense variety of topics and represent different types of texts, such as legal documents and charters, court proceedings, works of religious edification, pedagogical texts, medicine books, works on plants and on astronomy, etc. In total, the data being as they are, it is hard to imagine what a reliable sample in order to elaborate a partial "gold standard" could possibly look like. For the same reasons, building specialized tools, e.g. by creating an AN tagger lexicon, was clearly not feasible.

Instead, we had to work with existing resources and tools. But what started out as the second best option eventually turned out to be a very effective low-cost approach to our data, especially because the performance increased additionally after applying a layer of normalisation to our data prior to tagging. And since we normalised to a contemporary Medieval language, i.e., OF, our tagset did not need to be adapted, thereby allowing straight-forward comparisons across both languages. This work is meant to be of mutual benefit to the AND project (and eventual follow-up projects) and to the BASICS project on medieval language contact likewise. This contribution presents a snapshot of the work in progress, and we will refer to this stage as version 0.2 of the annotated corpus. In what follows, we describe the steps leading up to this version.

The first step consisted in ignoring the non-French passages[6] in the corpus. We did so for two reasons, firstly because they would have hampered the function of our analysis tools, which were trained for Old French. And secondly, because non-French passages and editorial notes are of no particular interest for the BASICS project. The XML markup of the texts could be used to identify most of the non-French passages, but some of them remained in the data and could not be dealt with manually at this point.

The second step was the segmentation, i.e. word form tokenisation and sentence splitting. On both the lexical and the syntactic levels this task is not trivial, but it does have a strong influence on the accuracy of automatic annotation. Since we use machine-learning tools for both tagging and parsing, the best results are achieved if word tokenisation and sentence splitting matches as closely as possible the texts the tools were trained on. The part-of-speech tagger (*TreeTagger*, Schmid 1994) uses parameters containing a lexicon of graphical

6 Non-French passages were, for example, Latin sentences in the psalters, and English and Latin paragraphs in macharonic texts.

forms most of which are associated with a lemma, so matching the input forms with the lexicon is important not only for the prediction of part-of-speech tags, but also for successful lemmatisation. Some of these tokenisation issues will be explained in more detail in the following section.

The accuracy of syntactic parsing depends quite heavily on the correct prediction of part-of-speech tags (more than on lemmatisation: in fact, lemmatisation had not significantly improved parser accuracy in previous tests with Old French; see Stein 2014), so word form tokenisation is also relevant for parsing. Moreover, since the main task of the parser is to predict the structure of a "sentence" (or at least of syntactic units defined as the relevant segments for parsing), the units of the input (the ANdb) should ideally follow the sentence definition of the training corpus (the SRCMF, Stein & Prévost 2013). However, this would have meant manually applying the SRCMF guidelines for sentence segmentation to the ANdb, which was not feasible at this stage of the project. In SRCMF, the unit "sentence" is defined minimally, as a structure containing no more than one main verb (which entails for example that coordinated main clauses are separated). Previous tests had shown that a dependency parser encounters fewer problems when input units are too long than when they are too short. Since verse texts contain many lines that are only parts of sentences, often lacking a verb, we decided not to use lines as an input unit, but to apply the same principles as for prose texts, i.e., we defined the sentence boundaries based on the punctuation marks inserted by the editors of the texts. Compared to the SRCMF principles, this often results in units that are larger than a SRCMF "sentence", e.g. enumerations containing main verbs or coordinations of main clauses (which were separated in SRCMF, according to the guidelines on http://srcmf.org). Since the parser, trained on SRCMF, has never seen coordinated predicates on the level of the main clause, it reacts by predicting for one of the coordinated structures a seemingly arbitrary category, for example "SjPer" (personal Subject) as in (1):

(1) [Confessïon desfait [**SjPer** et runt [Obj Trestots les liens
confession undoes and cuts all the bands
[ModA ke pecchez fount]]]]
which sins make
'Confession undoes and cuts all the relations that sins create. (all1237cors78)

However, the internal structures of both sentences are parsed correctly, which means that for syntactic queries that target structures other than coordination the analysis is acceptable. Thus, defining larger sentence units is the preferred choice, since it avoids the risk of producing units too small for the parser to analyse.

After the pre-processing the original files of the ANdb (including the "normalisation" described in section 2, our corpus (as of version 0.2) contained 3,111,982 tokenised graphical forms in 148,353 "sentences". After the segmentation procedure punctuation marks were deleted (again because the OF tools were trained on texts without punctuation marks). They are not included in the count of tokenised forms.

2 Normalisation of Anglo-Norman

2.1 Why normalisation matters

Due to spelling anomalies and to certain decisions on behalf of the scientific editors of the texts we are dealing with, queries cannot reveal all relevant hits. For example, querying the ANdb in its first, non-"normalised" version for *enportent* 'they carry away' yields three hits, among others:

(2) Dampnedeu les maudit S'il enportent un dener.
God them curses if-they carry-away one dime.
'God will curse them if they carry away one (single) dime'.
(alexander, 4/4 12th ct., v607)

But there is (at least) one more, which is:

(3) preignent lour blee et lenportent
take.3.PL their wheat and it_carry-away.3.PL
'They take their wheat and carry it away' (1419, Liber Albus, 783)

This fourth occurrence is not found because the editor chose to not intervene on agglutinated articles and pronouns, and did not separate the article from the verb with an apostrophe. As a consequence, to get an exhaustive list of occurrences of *enporter* in the AN texts, the query would have to match not only all the forms of the verb, but also all the possible kinds of agglutinated articles and pronouns, such as *d', s', m', l', n', c'*, etc. This is rather inconvenient and error-prone. Because of situations like these, we opted for "normalising" the texts prior to annotation. Normalisation has been previously applied to other historical languages, namely to Early Modern English (Rayson et al. 2007), Middle High German (Dipper 2010) and Early Modern German (Scheible et al. 2011) as well as to the ARCHER-texts (Hundt, Schneider, and Oppliger 2016), who all report a considerable increase of tagger-accuracy on normalized data (10%). In our case, recognition (i.e. the number of tokens matched in the tagger-lexicon) improves by 40 % on normalised

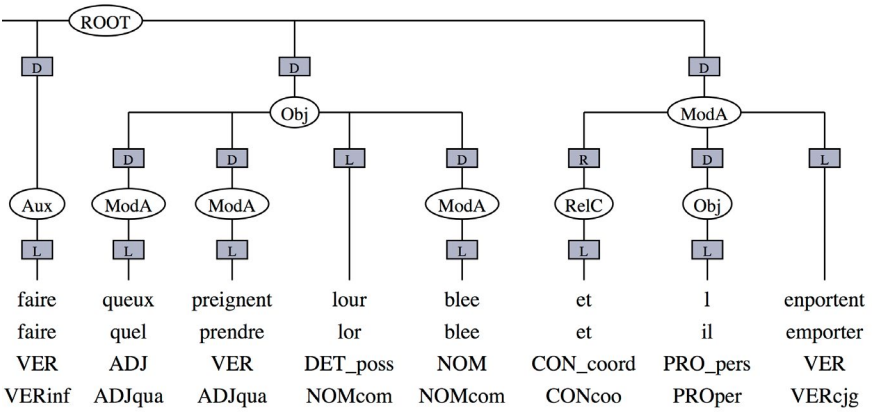


Figure 1: AN text data base, partial parse tree from albu783 (1225).

data. Unlike Dipper, and in line with Scheible et al., our approach does not involve retraining of the tagger. However, in contrast to Scheible et al, we do not intervene manually. Instead, we use an automated rule-based procedure and control the output of each single rule in order to prevent errors or over-generalisations. Also, in contrast to Rayson et al. and Scheible et al., we do not normalise to a modern standard and therefore do not have to intervene on the tagset itself, thereby maintaining straightforward comparability across both corpora.

Our goal is thus a POS-tagged and syntactically annotated version of the ANdb that allows us to retrieve, for example, not only the occurrence from the *Liber Albus* quoted in (3), just by searching for *enportent*, but ideally also the occurrence of "l" as a direct pronoun that is governed by the verb. The structure in Figure 1 is an example occurrence (see section 3 for an explanation of the dependency graphs).

Finally, we would like to point out that we did not normalise the text in the Lachmannian sense of the word. Rather, we calculated normalised forms in order to facilitate the identification of a given graphical form of the text in the tagger lexicon. If the generated form was successfully identified in the lexicon, the algorithm substituted the original form with the normalised form and did all further calculations on the basis of the generated form. But in the end, the generated form was, in turn, replaced by the original form, and all modifications that a given form had undergone remain invisible on the surface. In other words: no signs of intervention remain in the output.[7]

7 There is one change that nevertheless remains visible in the output, which is the separation of agglutinated forms. But this is state of the art in terms of "toilette du texte" (Foulet/Speer 1979, Lepage 2001 and École nationale des Chartes 2001). If a given text

2.2 Steps in normalisation

2.2.1 Preparatory measures

Of the 2,804,409 French tokens contained in the pre-processed version of the ANHub text-database, roughly two thirds (67.7 %) were matched by the Old French *TreeTagger* dictionary. Since the dictionary is all lower case, lower-casing all tokens raised successful identification to three quarters (75.09 %). At this point, we used a script to separate punctuation marks from words, because, given the fact that the dictionary contains only non-punctuated lemmata, tokens including punctuation marks would not have been retrieved. Tokenising increased the number of tokens to 3,439,145, four fifths of which were recognised (81.01 %).

The subsequent treatment described in the next sections below is based on this tokenised version of the ANdb (version 0.2). All the items that the tools could not identify in the dictionary at this point were submitted to further treatment.

2.2.2 Mechanical measures

In this step, we developed context-dependent rules for graphical normalisation. A graphical form that could not be matched in the first place underwent a series of successive context-sensitive modifications. However, while e.g. *lamour* is usefully converted into *l'amour*, *malaise* should be maintained as *malaise.* Therefore, each of these modifications was independently evaluated for success. This was easier with regard to pure graphical phenomena, such as e.g. the graphemes *y, k,* and *z*. In many cases, these graphemes are not used primarily for phonological reasons, but merely represent a variant spelling for *i, q(u)* and *(t)s* respectively. In these cases, they are fairly easy to replace, but the context has to be accounted for. E.g. *ey* equals *oi* in *neyent* 'nothing', *ai* in *faim* 'hunger', *eoi* in *receoit* 'he receives', *rey* in *derein* 'the last one', etc. In the end, y-rules were successfully applied in 23,164 cases.

The next step was to take into account regular phonetical features of AN, such as e.g. the spelling *ou* for *o*, or *om* for *ons*. These cases are of particular importance when it comes to suffixes, because, if a token such as *allom* 'we go, walk' was not recognised as a verb because of its ending in *om* instead of *ons*, the syntactic parser is also likely to fail at this point. Similarly, if *ioun* at the end of a word was converted into *-ion* and subsequently recognised as a word with a

lacks this kind of separation in the printed edition, it is no hallmark of the historical text, but the (modern) editors' choice. As such, there is no historical importance to it.

nominal suffix—that is, a noun—this benefits the part-of-speech and the syntactic analysis, even if the word is not in the tagger lexicon.

2.2.3 Additional measures

In addition to the substitutions described above, we had to intervene at two more points, one of them being proper names and the other agglutinated consonants. The latter keep the tagger from recognising the word even if the word itself is listed in the very same spelling in the dictionary, and in the case of the former, normalisation is not applicable.

Due to the nature of the texts included in the ANdb, many of which are legal documents and court proceedings, there are a considerable amount of proper names for both persons and places. In order to tag these adequately, we had to extract them and add them to the tagger-lexicon.[8]

We adopted two different approaches for extraction. Firstly, we collected all capitalised words from the file of unknown words generated by the tagger. In order to distinguish capitalised sentence initials from proper names, we sorted these forms by frequency, on the hypothesis that conjunctions etc., which might appear with a capitalised initial at the beginning of a sentence should occur as such more than once. In addition to that, sheer word-length allowed us to sort out a good deal of capitalised conjunctions, in contrast to proper names, which tend to be longer. This procedure allowed us to add 612 proper names to the tagger-lexicon and then re-train the tagger. Having again selected all capitalised forms from the new unknown-file, we sorted alphabetically, this time by the end of the words. This procedure helped us to detect the most frequent suffixes, such as e.g. *-fred* in person names or *-borough* and *-thorp* etc. in place-names. In the next step, we automatically extracted all forms ending in the 86 most frequent suffixes (down to frequency-rank 8) and added 2,473 additional proper names ending in the respective suffixes to the tagger lexicon. In total, we thus added 3085 additional entries. This step raised the overall recognition rate by roughly 1%.

The other approach dealt with agglutinated consonants. In AN, as in OF in general, words beginning with a vowel can combine with a consonant such as *c, d, l, m, n, q, qu, s, t* and the respective capital letters. A sequence like *l'article* would thus read *larticle* in the manuscript, and it would not be tagged correctly unless the agglutinated *l* was separated from the main word. On the other hand, this case has to be carefully distinguished from *malaise*, which should not be split into *m'alaise*.

8 Most taggers also exploit the „suffixes" of words to predict the category; the TreeTagger also applies such an algorithm to unknown word forms.

Therefore, we included a routine that checks unknown words for the initial sequence of "agglutinate consonant + vowel". If a word matches this pattern, the algorithm experimentally splits off the consonant and resubmits both elements to the recognition-procedure, this time analysing both parts independently, and writing successfully treated forms into an extra file. This output was checked manually. False recognitions, such as *d'estrece* 'narrowness' built from *destrece* 'hardship, affliction' (hypothetical example) were collected in a separate file. The routine then checked this file before proceeding to the treatment of possibly agglutinated forms. Doing this, the number of tokens was raised to 3 448 633, and, based on this new number, the rate of forms recognized by the tagger is at 92.94 %.

As one can see, it is indeed possible, and even at very low cost, to raise the rate of recognition by some 40%. One way to achieve this is by preprocessing the texts through "normalisation". By applying the procedures described above, we were able to normalise about 164,000 tokens equalling 39 000 types, with maximum token frequencies of up to 1,340 for forms of *estre* 'be' (1,340 *sunt*, normalised to *sont*, 3.pl.ind., and 1,328 *seyt*, normalised to *soit*, 3.sg.subj.).

The other way to raise recognition is by adapting the tagger and its lexicon, as they had originally been trained on continental French data, in order to cope with AN texts. Overall, "normalization" increased the rate of AN forms that are successfully identified in an OF tagger-lexicon by 25 percent points, from 67.7 % to 92.94 %—a step which will be crucial for the subsequent syntactic analysis.

3 Automatic syntactic analysis

3.1 Old French corpus annotation applied to Anglo-Norman

After the "normalisation" of the data described in the previous section, we applied a part-of-speech tagger and a dependency parser to the ANdb. Both were previously trained on Old French text corpora.

For part-of-speech annotation and lemmatisation, we used the *TreeTagger* with parameters for Old French. The tagger was trained on the *Nouveau Corpus d'Amsterdam* (Kunstmann/Stein 2007) and used a lexicon with form-tag-lemma triples that were extracted from various Old French resources[9]. This lexicon was identical to the one that was used for verifying the output of the normalisation rules described in section 2.3.

9 The training of *TreeTagger* and the lexical resources are described in Stein (2007). The lexical resources are freely available as *FROLEX*, see https://github.com/sheiden/Medieval-French-Language-Toolkit.

For dependency annotation we decided to use the *mate tools*[10] *joint transition-based parser* (Bohnet et al. 2013) for joint part-of-speech tagging and parsing. The parser was trained on the dependency annotation of the *Syntactic Reference Corpus of Medieval French* (SRCMF, Prévost/Stein 2013). The training corpus extracted from SRCMF contained 12 texts or text samples, written between 1000 and 1300, and containing 242,946 word tokens (23,818 types). Punctuation was not present (since modern punctuation appears only in modern transcriptions), and orthographical variation was considerable: the type-token ratio was more than twice as high (0.099) than in average Modern French texts (0.05), with the obvious negative consequences for the precision of part-of-speech tagging. The syntactic categories in the training corpus were a slightly simplified set of the SRCMF categories (see the documentation on the corpus web site http://srcmf.org).

The *joint transition-based parser* was chosen because it performed slightly better than the *mate tools* graph-based parser (Bohnet 2010) we had trained on the same corpus. Accuracy scores were better both for part-of-speech tags and labeled dependency attachment. More importantly, the joint transition-based parser also attained a higher score of exact sentence matches (i.e. where all the dependencies and categories in a sentence were analysed correctly) on our Old French evaluation corpus. The training procedure and the two *mate tools* parsers are described in greater detail in Stein (2016).

Concerning the results of this parser as applied to the Anglo-Norman texts, our expectations are not high. With a labeled attachment score of 85.96 % and a score of 47.59 % for exact sentence matches on the evaluation part of the SRCMF (i.e. a corpus containing the same text types), it is clear that the uncorrected output will present a considerable number of errors. Due to the particular characteristics and the heterogeneity of the AN texts described above, the parser is bound to perform worse, and we expect only very short sentences to be parsed correctly. An example for such a short sentence with correct analysis is given in (4), where according to the SRCMF markup, "Cmpl" is the indirect object, "RelNC" a non-coordinating relator (here: preposition), "Obj" the direct object, and "ModA" a modifier (including also determiners):

(4) A lui comand la meie vie
To him command.1.SG the my life
'I command my life to him.' (1250resu)

The output format of the parser is the CoNLL 2009 tabular format (defined on the CoNLL 2009 shared task web site, see http://www.conll.org). For the sake of clarity, Figure 2 shows a simplified CoNLL format representing only selected

10 https://code.google.com/archive/p/mate-tools/

columns: word number, form, lemma(s), TreeTagger POS, parser POS, morphological features, head attachment, and dependency relation. The last two columns encode the dependency structure. For example, "0" marks the verb *comand* as being the root node. "3" attaches *lui* (word no. 2) to *comand* (word no. 3), and the dependency relation is "Cmpl", i.e. indirect object. Likewise, "2" attaches *A* (word no. 1) to *lui* as a "RelNC" (non-coordinating relator), and so forth.

```
1  A       a          PRE       PRE        _                          2  RelNC
2  lui     il|loi     PRO:pers  PROper     G=masc|N=sg|C=obj|P=3      3  Cmpl
3  comand  comander   VER       VERcjg     N=sg|P=3                   0  ROOT
4  la      le         DET:def   DETdef     G=femi|N=sg|C=obj          6  ModA
5  meie    mien       ADJ:poss  ADJqua     G=femi|N=sg|C=obj          6  ModA
6  vie     vie|voie   NOM       NOMcom     G=femi|N=sg|C=obj          3  Obj
```

Figure 2: CoNLL format (simplified) for sentence (4).

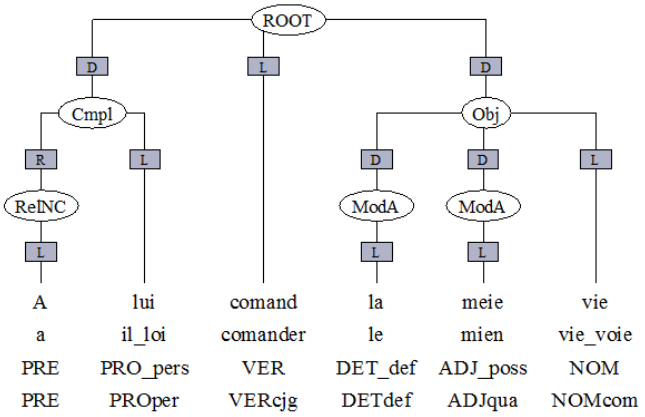


Figure 3: TigerSearch graph for sentence (4).

The CoNLL format can be used directly with some query tools like *Icarus* (Gärtner 2010). However, in the next section we use *TigerSearch* queries (Lezius 2002), since this is the default distribution format of the Old French SRCMF corpus. We therefore converted the CoNLL output of the parser into TigerXML. In Figure 3, the structure is shown as represented in the *TigerSearch tool*. In order to represent the SRCMF dependency graphs in *TigerSearch* (which was primarily designed to represent constituency structures), we distinguish between two kinds of relations (arcs): the default relation is dependency, labelled with a "D", whereas "L" marks the unique lexeme that governs the structure and

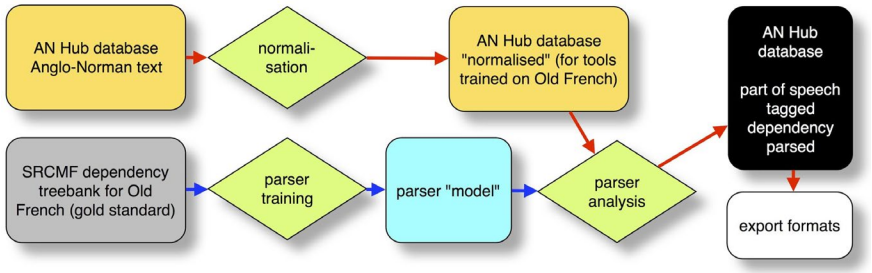


Figure 4: Annotation of the Anglo-Norman text database.

would figure as the top node of the structure in a traditional dependency graph à la Tesnière. For example, the main verb *comand* is attached to the root node by the "L" relation.

The complete workflow of the annotation is resumed in the flow chart shown in Figure 4. In the next section we will discuss the usability of the output.

3.2 Usability of unsupervised parsing

3.2.1 A case study

Since there is no gold standard corpus for AN, we cannot provide a quantitative assessment of the annotated ANdb. The goal of the cooperation between the BASICS and the AND projects was meant to be a feasibility study rather than an annotation project in its own right. We decided to use the annotated output for a research question that was relevant for the BASICS project anyway: the variation between direct and indirect objects that was observed in AN e.g. by Ingham (2010). These cases of variation are, for example, relevant for the development of passive structures in the medieval contact situation between English and (Anglo-)French. As pointed out in Stein and Trips (accepted), language contact with OF and AN may have attributed to the rise of the recipient passive (e.g., in ModE, *She was given the book*), since in Middle English corpora, the first occurrences of the recipient passive appear predominantly with verbs of French origin. So our analysis bears on OF ditransitive constructions. Just as in Modern French, continental OF had a dative goal (or recipient) phrase, i.e. a prepositional phrase governed by *a* (ModF à), for example with the verb *demander* ‚ask', as in sentence (5):

(5) et demande a Lancelot quele aventure l' a ilec amené
and asks.3.SG to Lancelot which adventure him has here brought
'and asks Lancelot which adventure brought him here (SRCMF, qgraal)'

One of the hypotheses we wanted to verify using the annotated ANdb was that the argument structures of AN ditransitive verbs was different from the (continental) OF structures, showing variation between indirect and direct objects. In order to do so, we needed to extract these verbs in specific constructions from the corpus. In the following subsections, we describe the relevant queries step by step, from the word level to the syntactic level, and discuss the advantages and problems we encountered in the annotation.

3.2.2 Lemmatisation

At word level, the first step was the selection of a representative sample of clause-taking verbs. We used the lemmatisation introduced by *TreeTagger* to query for eight such verbs, i.e. *assëurer, demander, certefiier, comander, garnier, informer, prier, vëer.*[11] We manually checked the precision of the results. It was generally satisfying, i.e. the result did not contain many forms not matching these verbs, except for some prefixed forms (forms of *deprier* instead of *prier*). The recall (i.e. the relation between the extracted forms and those which could have been maximally extracted) can only be estimated. We again verified manually and found that recall was not lower than if we had performed a search targeted at inflected forms, using regular expressions. This is probably due to the fact that AN graphical variants are fairly unpredictable (as was shown in section 2). Nevertheless, by querying the lemmas we found a number of graphical forms that would have been hard to guess, as for example *Nos te praeiam* (*nous te prions*, 1.PL., 'we pray you'). And queries aiming at particular verb classes (which often have many more than the eight members we selected for our example) would be extremely laborious if lemmatisation was not present in the annotation. So we can conclude that unsupervised lemmatisation, even if it is only partial and may contain errors, is indeed useful.

11 The TigerSearch query specified the following lemmas:
`[lemma=/.*(assëurer|demander|certefiier|comander|garnier|informer|prier|vëer).*/]`

3.2.3 Ditransitive constructions

The second task was to narrow the output down to ditransitive constructions. This step *requires* syntactic annotation (or manual analysis, which is not at issue here). Querying ditransitive constructions using only part-of-speech annotation is extremely laborious. It requires a combination of several subsequent queries, and would probably lead to low precision and recall values. This is due to the variable position of each of the arguments in the clause, the graphical variants, and the syntactic ambiguities, where the first two factors affect recall, and the latter affects precision (not every prepositional phrase is an indirect object, etc.). The SRCMF grammar model reproduced by the parser allows extraction of ditransitive constructions with a single query, e.g. in *TigerSearch*. This query[12] finds a total of 365 sentences (265 with a verb form of *demander*, 168 with *comander*, 26 with *prier*, 26 with *vëer*, etc.).

3.2.4 Clitics vs full lexical arguments

In the next step, we were interested in the various forms of argument realisation. Since AN texts sometimes show inconsistencies in the use of clitics (Old French distinguishes between accusative and dative pronouns), we are interested in the different combinations of clitic and full lexical argument realisation. Clitics appear preverbally, i.e. at a position that is normally different from the (generally postverbal) position of full nominal arguments. Again, clitics are very difficult to retrieve unambiguously: their graphical forms are extremely variable, and they are often homographs of other grammatical morphemes like articles (*le*, *li*, etc.). Using the syntactic annotation, we retrieved clitics by combining POS tag ("PRO") and node "arity", the latter being "1", since clitics do not govern other nodes. The *TigerSearch* query given in footnote[13] is meant to serve as an example: it extracts only the occurrences where both the direct and indirect object are clitics, which is in fact quite rare. In our project, we are rather interested in cases that are analogous to *She commanded him to leave*, in order to find out if the goal argument (*him*) is an accusative or a dative clitic in the Anglo-Norman construction. So, one of the arguments needs to be specified as being clausal. We

12
```
#s:[type=/V.*/]
   & #s > #v:[< list of lemmas, as needed>]
   & #s > #a1:[cat="Obj"]
   & #s > #a2:[cat="Cmpl"]
```

13
```
[lines 1-4 identical to first query]
   & arity(#a1,1) & #a1 > #acc:[pos=/PRO.*/]
   & arity(#a2,1) & #a2 > #dat:[pos=/PRO.*/]
```

further restrict our clitic query to third-person forms beginning with *l* (since first and second person do not distinguish between accusative and dative). Finally, the goal argument, which in continental Old French normally has dative case, is specified as direct object ("Obj", i.e. accusative).[14]

We applied this query to the SRCMF corpus as well as to the ANdb. In SRCMF we obtained only one result (from the *Chanson de Roland*, an early Anglo-Norman text):

(6) Par penitence les cumandet a ferir
By regret them.ACC commanded to strike
'He regretfully commanded them to strike.' (roland-pb: 100-lb1138)

In the ANdb, the query retrieved ten occurrences, which could confirm that the variation between dative and accusative clitics in clause-taking ditransitives is indeed characteristic of Anglo-Norman. The precision, however, was low: in addition to the ten valid examples we retrieved many erroneous hits where the parser annotated the wrong structure. A typical error is the non-recognition of dislocations, as in example (7):

(7) donets moy grace qe jeo le voille et jeo soie si treshumble pacient come
le mestier le demande a receivoir bonement les cures
the profession$_i$ it.ACC$_i$ requires [to accept well the cures]$_i$
'The profession requires it to receive the treatments willingly.' (1354seyn2374)

Even for the human reader, it is not an easy task to detect that *le* preceding *demande* is not the goal argument here, but a cataphoric clitic that doubles the right-dislocated clausal complement, i.e. *a receivoir bonement les cures* (both are co-indexed with *i* in the glossed example). So in fact, this example is not an instance of ditransitive *demander*. Again, if we wanted to measure the recall of the query we would have to check for missed occurrences, using a series of word form and POS-based queries.

The last variant we discuss here is the case where the goal argument is a full NP and the clause is the theme. Again we want to find out if the goal argument is a direct or indirect (prepositional) argument, i.e. "Obj" or "Cmpl" in terms of SRCMF categories (analogous to English constructions like *She commanded (the*

14
```
#s:[type=/V.*/]
   & #s > #v:[<list of lemmas, as needed>/]
   & #s > #a1:[cat=/Obj|Cmpl/ & type=/V.*/]
   & #s > #a2:[cat="Obj"]
   & arity(#a2,1) & #a2 > #dat:[word=/l.*/ & pos=/PRO.*/]
```

knight|to the knight) to leave). In the query given in the footnote[15], we defined the goal argument as non-verbal, specifying a minimal arity of 2 (thus eliminating clitics) and added a restriction for linear precedence (goal occurring before clause). Again, precision was low: the query produced noise due to parsing errors in complex sentences. A good result is example (8), whereas in example (9) the subject was wrongly parsed as a direct object:

(8) ... et demandent **les marchans** a avoir du maistre leurs denrees
... and ask.3.PL the merchants.ACC to have from-the master their goods
'and they ask the merchants to get their goods from the master'
(1310domg1769)

(9) Et comande le Rei qe les Viscontes ...
And comands the king.NOM that the viscounts ...
'and the king commands that the viscounts ...' (1275stat110)

3.2.5 Analysing grammatical variation

A particular problem arises when the corpus analysis targets grammatical variation. Variations like the one mentioned above, between accusative and dative clitics, are notoriously difficult to identify using machine-learning approaches. In our case, the variation is said to be typical for later AN. Since the parser was trained on the SRCMF texts, it cannot be expected to have encountered this kind of variation. Therefore, when the less frequent option of a particular instance of grammatical variation is encountered in the input data, this will create a conflict at the syntactic level. In the example (6), the clitic *les* is part-of-speech tagged as accusative, but it co-occurs with a verb that normally governs a dative complement (*comander*). It is rather unpredictable, at least for the linguistic user, if the parser will select the category, i.e. direct *vs* indirect object, that matches best the part-of-speech analysis or the valency of the verb. In our corpus, the *joint transition-based parser* seemed to be more strongly influenced by the part-of-speech information. That means that the linguistic perspective, which describes this case as variation on the morphological level, cannot be translated directly into a

15
```
#s:[type=/V.*/]
& #s > #v:[< list of lemmas, as needed>]
& #s > #a1:[cat=/Obj|Cmpl/ & type=/V.*/]
& #s > #a2:[cat="Obj" & type!=/V.*/]
& arity(#a2,2,99) & #a2 .* #a1
```

query. Instead, the user has to anticipate the way the parser analyses these cases when formulating their query. Examples like (6) can only be retrieved by a query that specifies the goal argument as direct object ("Obj") on the syntactic level or underspecifies the syntactic category.

4 Conclusion

The goal of this contribution was to demonstrate how linguistic tools that were previously trained on other varieties of a medieval language can be applied to a specific variety of this language using „normalisation" techniques. In our case, the medieval language was Old French (OF), and the new corpus was the Anglo-Norman text database (ANdb). Since graphical conventions in Anglo-Norman (AN) are quite different from those of continental OF, we normalized the AN texts before applying the computational-linguistic tools. We use "normalising" in the sense of adapting the AN forms to the continental OF spelling conventions as closely as possible. We used the OF lexicon contained in the parameters of *TreeTagger* to measure the score of normalised forms and showed how graphical normalisation, including the resolution of determiners that are agglutinated to nouns, improves the performance of the tools. We (partly) lemmatised the corpus using *TreeTagger*, and added dependency structures using the *mate-tools joint transition-based parser*. Since a gold standard corpus for Anglo-Norman does not exist, we were unable to calculate accuracy scores for these analyses. Instead, we evaluated the quality of the annotation from a linguistic point of view, searching for particular argument realisations of ditransitive verbs.

As expected, the major issue due to errors in the annotated version of the ANdb is low recall, and it is hardly measurable how many of the structures we queried were not successfully retrieved. We showed that, in some cases, a good feeling for the way the parser works is required to anticipate its analyses and to formulate the queries accordingly. This issue hampers the quantitative interpretation of the data. However, we also saw that parsing, albeit imperfect, allows us to make queries and extract occurrences for structures we could not have retrieved otherwise (at least not in acceptable time). Thus, even with medieval texts, the unsupervised use of computational tools, paired with a normalisation procedure that graphically adapts the novel text to the graphical conventions of the training corpus can help to extract relevant syntactic data and thus assist diachronic syntactic analysis. Especially with larger amounts of data (as in the case of the ANdb, containing over 3 million words) parsing, even with low accuracy, may be the only way to discover certain phenomena and to retrieve the relevant data.

References

Bohnet, Bernd. 2010. Top Accuracy and Fast Dependency Parsing is not a Contradiction. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, Beijing, China: Coling 2010 Organizing Committee, 89–97.

Bohnet, Bernd, Joakim Nivre, Igor Boguslavsky, Richárd Farkas, Filip Ginter and Jan Hajic. 2013. Joint Morphological and Syntactic Analysis for Richly Inflected Languages. *TACL 1*, 415–428.

Dipper, Stefanie. 2010. POS-tagging of historical language data: First experiments. In *Proceedings of the 10th Conference on Natural Language Processing (KONVENS-10)*, Saarbrücken.

Gärtner, Markus, Gregor Thiele, Wolfgang Seeker, Anders Björkelund and Jonas Kuhn. 2013. ICARUS – An Extensible Graphical Search Tool for Dependency Treebanks. *Proceedings of ACL 2013*.

Grant, Judith. 1978. *La passiun de seint Edmund.* Oxford: Blackwell.

Hundt, Marianne, Gerold Schneider, Rahel Oppliger 2016: Part-of-Speech in Historical Corpora: Tagger Evaluation and Ensemble Systems on ARCHER. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS-13)*, Bochum.

Hunt, Tony. 2004. *Le Chant des chanz.* London: Anglo-Norman Text Society.

Ingham, Richard. 2010. The Transmission of Later Anlo-Norman: Some Syntactic Evidence. In Richard Ingham (ed.), *The Anglo-Norman Language and its Contexts*, 164–182. Woodbridge: Boydell and Brewer.

Ingham, Richard. 2012. Middle English and Anglo-Norman in Contact. *Bulletin de l'Association des Médiévistes Anglicistes de l'Enseignement Supérieur* 81: 1–23.

Johnston, Ronald Carlyle. 1961. *Crusade and Death of Richard I.* Oxford: Blackwell.

Kunstmann, Pierre and Achim Stein. 2007. Le Nouveau Corpus d'Amsterdam. In Pierre Kunstmann and Achim Stein (eds.), *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*, 9–27. Stuttgart: Steiner.

Lezius, Wolfgang. 2002. *Ein Suchwerkzeug für syntaktisch annotierte Textkorpora (German).* Stuttgart: Institut für Maschinelle Sprachverarbeitung (IMS).

Marchello-Nizia, Christiane. 2009. Histoire interne du français: morphosyntaxe et syntaxe. In Gerhard Ernst, Martin-Dietrich Gleßgen, Christian Schmitt and Wolfgang Schweickard (ed.), *Romanische Sprachgeschichte. Ein internationales Handbuch zur Geschichte der romanischen Sprachen und ihrer Erforschung, Teilband 3*, 2926–2947. Berlin, New York: de Gruyter.

Prévost, Sophie and Achim Stein. 2013. *Syntactic Reference Corpus of Medieval French (SRCMF).* Lyon/Stuttgart: ENS de Lyon; Lattice, Paris; Universität Stuttgart.

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper and Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS

tagger on Early Modern English corpora. In *Proceedings of the Corpus Linguistics Conference 2007*. Birmingham: University of Birmingham.

Rothwell, William and David Trotter. 2005. *Anglo-Norman Dictionary 2. Online Version.* London: MHR.

Scheible, Silke, Richard J. Whitt, Martin Durrell and Paul Bennett. 2011. Evaluating an 'off-the-shelf' POS-tagger on Early Modern German text. In *LaTeCH '11 Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, 19–23.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging using Decision Trees. In Daniel Jones (ed.), *Proceedings of the International Conference on New Methods in Language Processing (NeMLaP'94), Manchester, September 1994*, 44–49. Manchester: UMIST.

Stein, Achim. 2014. Parsing Heterogeneous Corpora with a Rich Dependency Grammar. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), 26.–31.5.2014*, Reykjavik, Iceland: European Language Resources Association (ELRA).

Stein, Achim. 2016. Old French Dependency Parsing: Results of Two Parsers Analysed from a Linguistic Point of View. In Nicoletta Calzolari et al. (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 23.–28.5.2016*, Portoroz, Slovenia: European Language Resources Association (ELRA).

Stein, Achim, Carola Trips (accepted): A comparison of multi-genre and single-genre corpora in the context of contact-induced change. In Richard Whitt: *Diachronic corpora, genre and language change*. Amsterdam, Philadelphia: Benjamins.

Joanna Bilińska, Monika Kwiecień, Magdalena Derwojedowa

Microcorpus of Nineteenth-Century Polish

Abstract In the paper, a 1M word corpus of Polish texts from the period 1830–1918 is described. The corpus was compiled to provide diversified linguistic data for morphological analysis, however several tests proved that it can be used as a versatile resource to identify various linguistic phenomena and trace their dynamics in regard to inflection, spelling or even syntax. It is divided into five equal subcorpora to provide stylistic variety: scientific texts for general public, news, feuilletons, fiction and drama. In order to conduct morphological analysis an analyzer made for contemporary texts was adapted, which can, therefore, process word forms that differ from contemporary inflection and spelling. In the paper, several experiments made with the use of the corpus are discussed.

Keywords Morphological analysis, spelling, 19th century Polish, corpus

1 Introduction

The aim of this paper is to present a 1M word corpus of Polish texts from the period 1830–1918, available as text samples and metadata files (http://www.f19.uw.edu.pl/download/korpus-f19-v1-0/).[1] A browsable version, using the Polish National Corpus Poliqarp engine (Przepiórkowski et al. 2012), is available at https://szukajwslownikach.uw.edu.pl.[2] Originally the corpus was compiled to deliver as much diversified data as possible for morphological analysis, any other research in diachrony or history of language being just an additional possibility facilitated by this project (cf. Derwojedowa et al. 2014a, b). The paper is organized as follows: in the first part we present the overall design of the corpus

1 This research was funded in the years 2013–2017 by the Polish National Science Centre grant DEC-2012/07/8/HS2/00570.

2 The instance of the corpus compiled for Poliqarp browser off-line is available at http://www.f19.uw.edu.pl/download/obraz-korpusu-1830-1918/.

(macrostructure), then we present the design of a sample (microstructure). In the next part, there is a discussion of some experiments conducted with the use of the corpus.

2 Corpus' structure

The corpus consists of 1000 samples of 1000 tokens each. The samples were divided equally into 5 subcorpora: scientific texts for the general public (1), news (2), feuilletons (3), fiction (4) and drama (5). This method differs from the choice of texts made for the Polish National Corpus (PNC, Przepiórkowski et al. 2012) and the corpus of Baroque-period Polish — KorBa (Gruszczyński et al. 2013; under construction), but such a division was well-tested on the small-scale corpus of *the Frequency Dictionary of Polish* (Kurcz et al. 1990). In the time span of our research, drama seems the best approximation of speech, but also the burgeoning vocabulary of emerging science, engineering and fast-changing social reality need to be taken into account. Tests such as cluster analysis and multidimensional scaling (Eder et al. 2013, cf. R-manual 2015) concluded that the texts are distinctively spread between styles (cf. Figure 1).

The overriding principle of the project was that first printed editions of texts written originally in Polish were included in the corpus. Some exceptions were applied in special cases (e.g. literary works first issued in episodes in a newspaper or a magazine; cf. Bilińska et al. 2016).

Most texts were acquired from digital libraries. Despite the rule of at least one sample per year in each subcorpus, the acquisition was a result of rather opportunistic guidelines: we searched sources with a text layer (e.g. plain text and/or layered djvu). If such a source was not available, which was the standard case for the earlier quarter of the period, we decided to OCR files in graphic formats (.jpg or .png).

The number of samples in a style for a given year never exceeds four. In the whole corpus, each year is represented by at least five samples but no more than twenty. The majority of years is represented by 10–13 samples with an average of 11 samples per year (cf. Figure 2).

3 Sampling the corpus

A sample comprises a couple of files: a fragment of continuous text, its metadata and a source graphic file (.png, .jpg, .djvu, .pdf, .tiff; cf. Figure 3). The excerpt — a proper text sample for research — is the most accurate representation of the source text. The footnotes, incomprehensible fragments, stage directions and

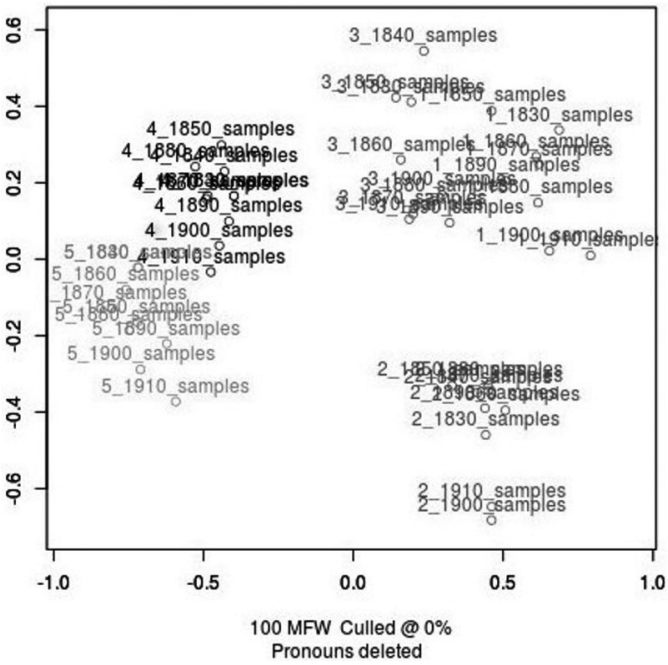


Figure 1. MDS grouping of styles (samples of each style merged by decades).

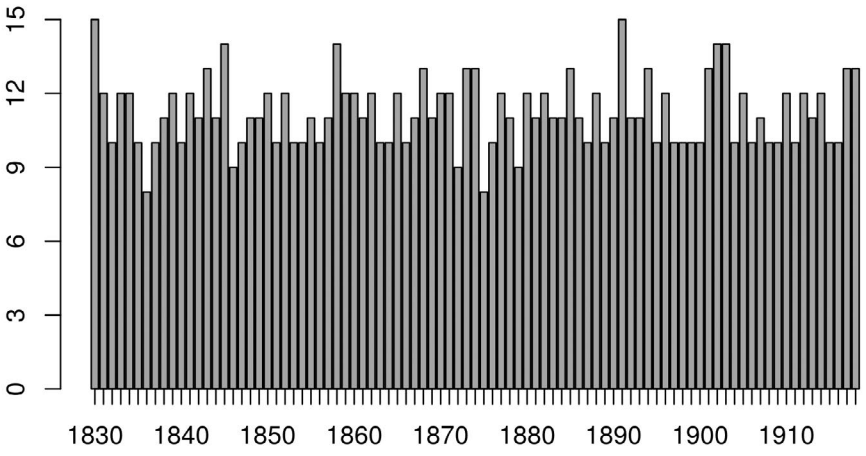


Figure 2. Number of samples per year.

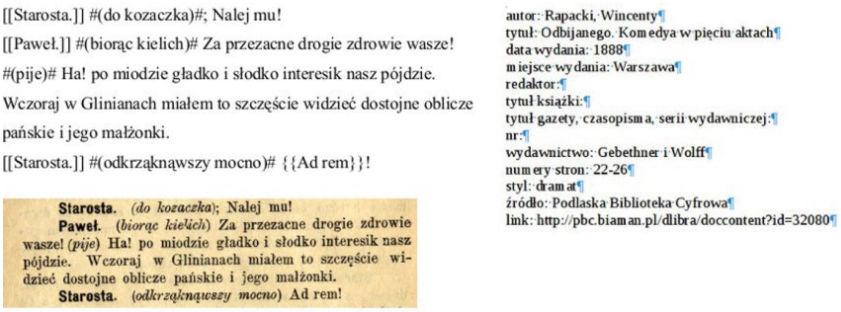


Figure 3. Text, metadata and source file of sample 1888_5.1.

small fragments in foreign languages, even misspellings were marked, but left unedited.

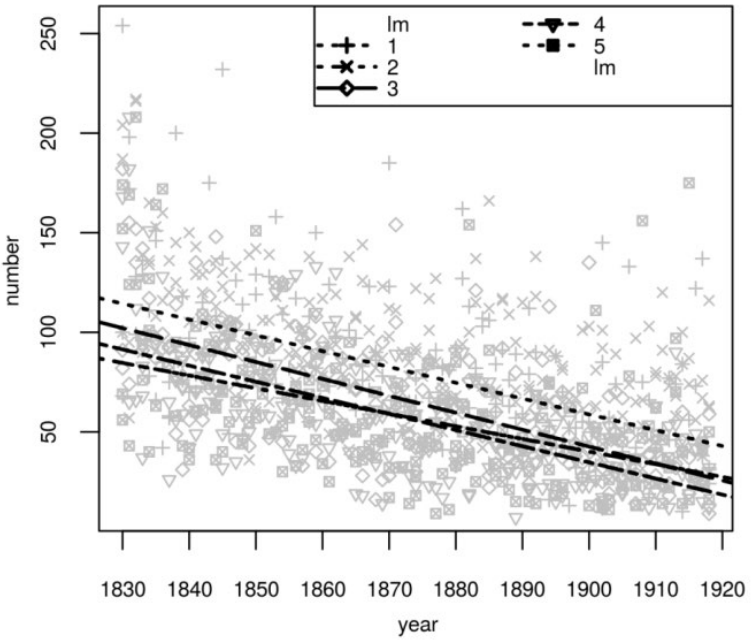
4 Diversity of the corpus

It is difficult to ascertain the exact number of authors without in-depth research (newspaper texts are often signed with initials or left unsigned; in the whole corpus there are 270 such samples), however there are circa 650 individual writers. Some are represented in more than one sample, but never in more than one style per year. In total there are 106 writers cited more than once.

Even though we struggled to create as diversified a collection of texts as possible, we did not select texts with respect to regional linguistic features. In effect, almost 2/3 of the texts were printed in Warsaw (almost 40 %), Lviv and Cracow. Together with texts issued in Paris, Vilnius, St. Petersburg and Leipzig they comprise almost 90 % of the corpus. The remaining 68 printing centers are represented several times and 39 of them – just once.

The majority of sources comes from big academic centers that undertook substantial projects of digitizing library archives. We used 43 such archives but 54 % of samples were excerpted from just three of them (Polish National Library on-line Polona, Warsaw University Digital Library, Digital Library of Wielkopolska).

The corpus is a resource of nineteenth-century Polish language indispensable for modifying a morphological analyzer in order to enhance its capabilities to analyze older texts. For this reason, we initially analyzed each sample and each subcorpus with an unmodified (i.e. trained on contemporary Polish) analyzer. Generally speaking, the number of unrecognized segments decreases with every newer sample and differs between circa 5 % and 15 % for a style and between 2 % to circa 25 % in case, respectively, of the best and the poorest sample in a given style (cf. Figure 4). The best results come from analyzing fiction, which can be



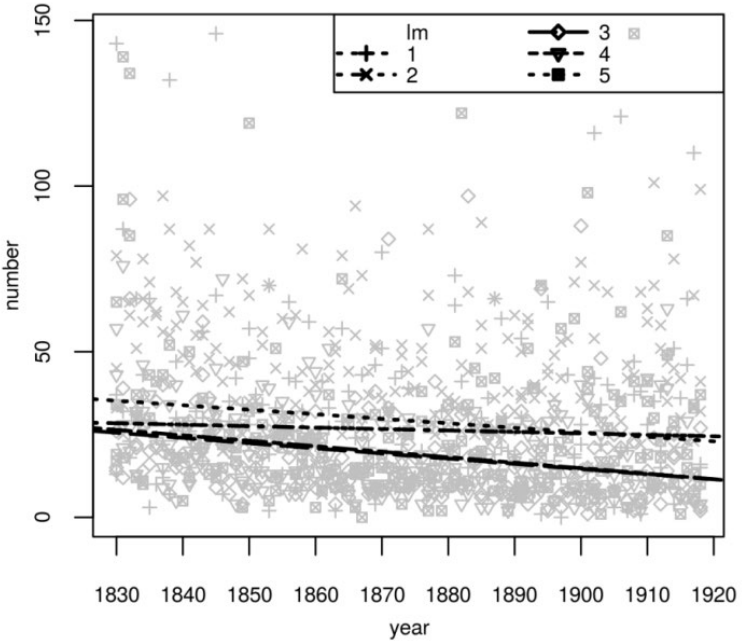


Figure 4. Unrecognized tokens in the 5 styles (1 is for science for general public, 2 —for press news, 3 — for feuilletons/journalism, 4 — for fiction, 5 — for drama), not modified analyzer. / Figure 5. Unrecognized tokens in the 5 styles (1 is for science for general public, 2 —for press news, 3 — for feuilletons/journalism, 4 — for fiction, 5 — for drama), modified analyzer (http://www.f19.uw.edu.pl/download-category/analizator/).

attributed to the fact that this type of language is mostly represented in dictionaries that constitute a base for any NLP device (cf. Saloni et al. 2015, Woliński 2014). For the same reason, the outcome of journalistic subcorpus' analysis is quite similar because this style is also included by lexicographers in the material base of their works. The poorest result comes from analyzing subcorpus of drama — in these texts there are, seen relatively, a large number of proper names, colloquial expressions, interjections etc.

5 Subcorpora

We will characterize each subcorpus in brief. The subcorpus containing scientific texts for the general public is comprised of samples excerpted from monographies, textbooks as well as scientific papers and popular science articles in the magazines. These were foremost the emerging Polish periodicals (written in Polish) aimed at popularizing current scientific achievements and discoveries especially in the life sciences. Magazines and books are almost equally represented.

In this subcorpus the morphological analysis gave results spanning from 1.3 % (sample from 1897) unrecognized segments to almost 25 % (sample from 1830). The reason for such a high percentage of unrecognized forms is not just spelling that was different from contemporary orthography but also foreign words in different stages of assimilation (e.g. *feldspat* 'feldspar'), technical terms and suggested Polish equivalents that were not accepted in the end (e.g. *błyszcz* 'stibnite (antimonite)').

The second subcorpus — containing short press texts — mainly consists of short relations from daily newspapers published in the biggest Polish cities. Apart from the daily press, newspapers issued twice or once a week and every two weeks were also considered, which was common for places with no daily press. The language of press notes did not differ from the language of scientific texts for the general public (2.3 % in the most recognizable sample, 25.3 % in the least recognizable one), however the main source of unidentified parts are different spelling or older forms of inflection.

The journalistic subcorpus includes texts published in newspapers, journals and books. The most characteristic feature of the style is the anonymity of texts — almost half of them are signed only by initials, a pseudonym or collective author. On the other hand, these excerpts are almost fully recognizable (0.9 % to 18.2 %, about 6 % on average), possibly because of the style's closeness to general language, the small number of foreign words and/or professional vocabulary.

The fiction subcorpus contains mainly samples of novels and stories. Seven samples of verse novels and epic poems may be treated as an exception, however they are typical for the earliest 25 years of the period. In metadata they

are marked as verse prose because this information may be useful for natural language processing. In novels and stories (mainly romances) from the earlier period there are many fragments in French, on the other hand there was very limited availability of prose texts at that time, so they cannot be replaced with other material. In the later samples, mainly older inflectional forms are not recognizable — the average is about 5.5 % with a range from 0.5 % to 20 %.

The drama subcorpus contains samples of different kinds of dramatic works — from the masterpieces of Polish playwriting to the libretti of operettas and vaudevilles. As stated before, the analysis of these texts gave the weakest results (1 % to 28 %, 8 % on average). It is most unlikely that these results can be improved because there are a lot of interjections, dialect words etc., even though the utmost care was taken to avoid texts with strong dialect, historical or parodic stylization.

6 Processes of linguistic change through the corpus' lens

In spite of its small size, the corpus may be used not only as a source of data for an analyzer but also as material for research on the linguistic processes of change in regard to inflection, spelling (cf. Derwojedowa et al. 2016) or, to some extent, syntax (it consists of more than 11,000 sentences). Clear distribution of texts between styles (cf. Figure 1) allows even the formulation of tentative hypotheses concerning the differences between the subcorpora. First of all, changes listed in grammar books (cf. Bajerowa 1986, 1992, Klemensiewicz 2001) were looked at more closely. There are about 20 features of that period that may be verified on small-scale datasets. Figures 6 to 9 provide some examples. Figure 6 presents an overall picture of the evolution of adjective endings in the nineteenth century—*-em(i)/ém(i)* and *-éj* made by Kopczyński (1817) and those inherited from earlier stages of Polish.

Figure 7 presents the dynamics of change in adjective endings in instrumental and locative singular and the instrumental plural of both masculine and neuter from late Middle Polish -ym(i)/-im(i) to nineteenth century. -ém(i)/-em(i) and earlier.

In Figure 8 contraction [ij]/[ij] → [j]/[i] in loanwords is shown. Bajerowa's (1986) claim that the process was almost finished at the time is generally right, however it seems that it is still active (even if only simmering) in a wider class of left context consonants than in her research. It can be clearly observed that mostly stem-syllables are affected, long syllables in the stem being rare (circa 70 wordforms of 30 lexemes in 650 wordforms altogether), with *austryj~* ('Austrian', presently *austri~*) being most frequent (cf. Figure 7).

When compared with the frequency of *Ross(y)ja*, *rossyjsk-* ('Russia', 'Russian' *Rosja*, *rosyjski*) and *Prussy*, *prussk-* ('Prussia', 'Prussian' *Prusy*, *pruski*) with respect

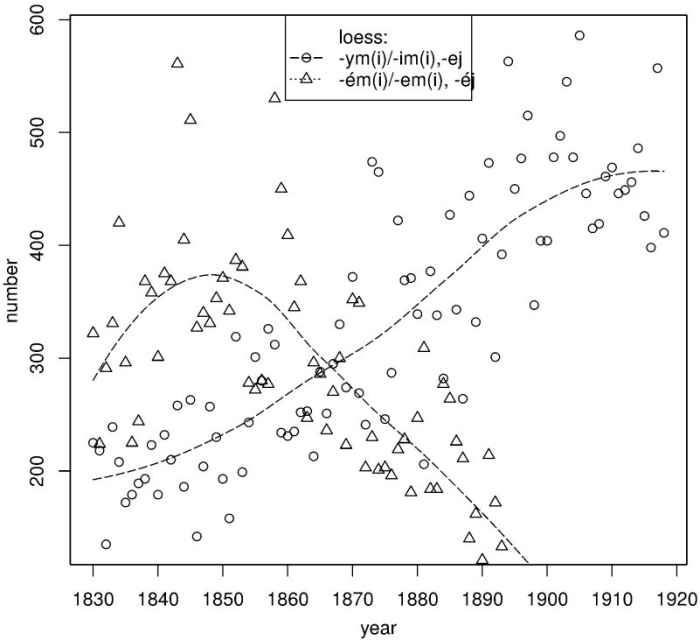


Figure 6. Innovative and historically developed endings of adjective-altering between 1830 and 1918. The dotted line represents innovative endings in total, i.e. any endings with é and e (loess = locally weighted scatterplot smoothing, cf. Cleveland et al. 1988).

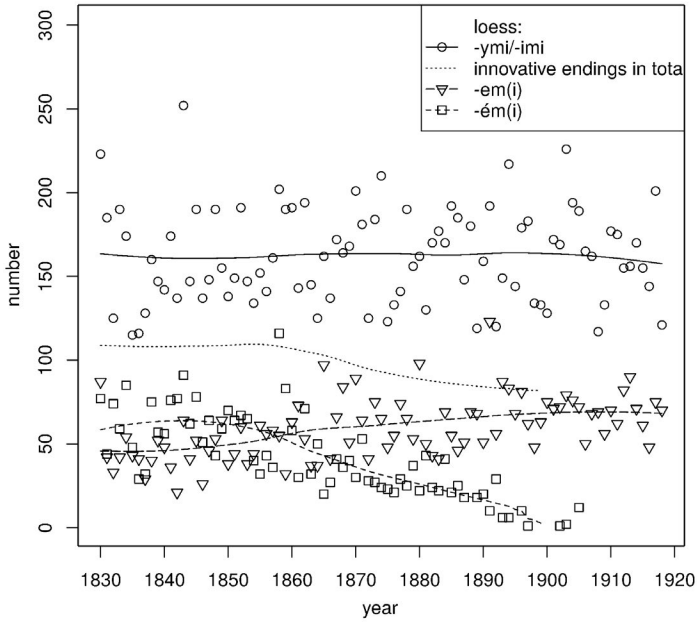


Figure 7. Innovative and inherited masculine and neuter endings of *adjectives* in instrumental and locative singular, instrumental plural, all genders.

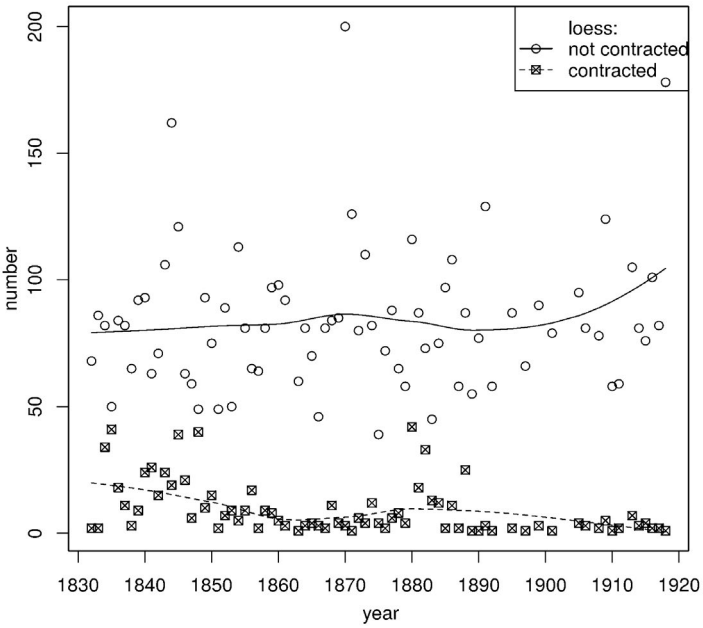


Figure 8. Words with contracted and uncontracted syllable [ɨj]/[ij].

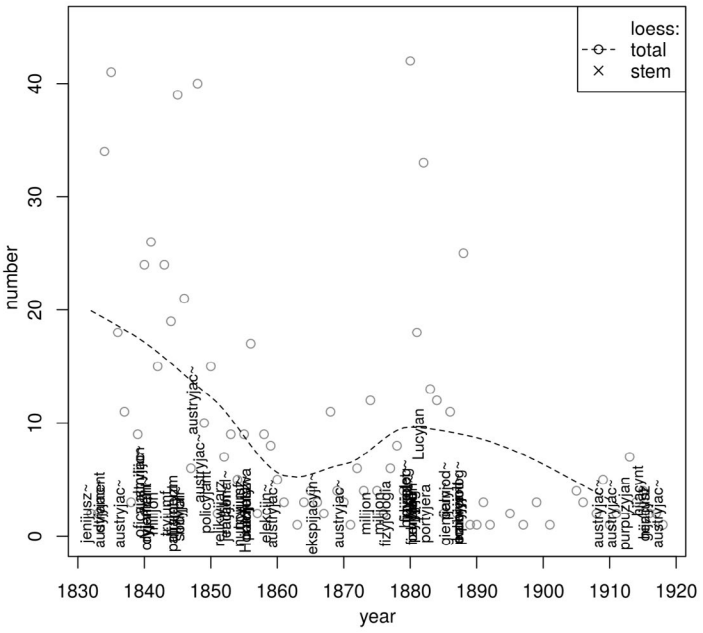


Figure 9. Uncontracted syllables in total, stems with not contracted syllables.

to the usage of doubled letters in loanwords, we clearly see that instead of processes, we rather observe lexical phenomena – all three are stems used in names of offices and institutions. Figure 9 shows the number of all contracted forms and points to individual uncontracted stems over the time span of 1830–1918.

The last example is the spelling of the (orthographic) string *ge* in loanwords. It is well attested that over time, the string became depalatalized in the period in question, being pronounced (and in consequence spelled) with *je*, *gie* and (innovative) *ge* by no other rule than according to a writer's belief or habit, e.g. spelling *jenerał* ('general') is almost three times more frequent than *generał*, no evidence of *gienerał*, whilst in the case of *geografia* and *jeografia* ('geography'), the spelling is exactly the opposite, with just one *gieografia*. The Dictionary of Polish by Niedźwiecki, Karłowicz and Kryński (1900–1927) quotes over 1,300 entries with *gie*, while there are less than 30 words with *gie* in the corpus. Some of them are lexical derivates (e.g. *Giermanie* 'Germans' and *giermański* 'German, adj'), and are present only in 5 % of samples. All others are spelled with an original *ge*.

7 Conclusion

Until now, neither a balanced, tagged and verified corpus of nineteenth century Polish nor an analyzer able to process older Polish texts have been available. Because of relatively small samples, the diversity of the corpus in many respects (places, authors, printed sources etc.) is quite satisfactory. Several tests passed on the corpus have proved that it can be used as a versatile resource to identify linguistic phenomena, trace their dynamics (cf. Figures 4–7) and turning points or to confront the emerging rules of orthography and good usage from the grammar handbooks with everyday practice. The corpus may be treated as an independent resource for research in inflection, morphonology and, to some extent, syntax. The considerable differentiation of samples makes it useful as an initial resource for research in new vocabulary and lexical changes as well.

References

Bajerowa, Irena. 1986. *Polski język ogólny XIX wieku. Stan i ewolucja.* T. I. *Ortografia, fonologia z fonetyką, morfonologia.* Katowice: Uniwersytet Śląski.

Bajerowa, Irena. 1992. *Polski język ogólny XIX wieku. Stan i ewolucja. T. II. Fleksja.* Katowice: Uniwersytet Śląski.

Bilińska, Joanna, Magdalena Derwojedowa, Monika Kwiecień and Witold Kieraś. 2016. Mikrokorpus polszczyzny 1830–1918. *Komunikacja Specjalistyczna*, in print.

Cleveland, William S. and Susan J. Devlin. 1988. Locally-Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* 83 (403): 596–610. doi:10.2307/2289282. JSTOR 2289282 Derwojedowa, Magdalena, Witold Kieraś, Danuta Skowrońska and Robert Wołosz. 2014a. Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych. *Polonica* XXXIV: 21–27.

Derwojedowa, Magdalena, Witold Kieraś, Danuta Skowrońska and Robert Wołosz. 2014b. Zasób leksykalny polszczyzny II połowy XIX wieku a możliwość automatycznej analizy morfologicznej tekstów z tego okresu. In Małgorzata Gębka-Wolak, Joanna Kamper-Warejko and Andrzej Moroz (eds.), *Leksyka języków słowiańskich w badaniach synchronicznych i diachronicznych*, 183–196. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Derwojedowa, Magdalena, Witold Kieraś, Joanna Bilińska and Monika Kwiecień. 2016. Dynamika zmian między reformami 1830–1918. *Język Polski* z. 1 XCVI: 24–35.

Eder, Maciej, Mike Kestemont and Jan Rybicki. *Stylometry with R: a suite of tools*, *leksykalny polszczyzny II poł. XIX wieku a możliwość automatycznej analizy morfologicznej* Digital Humanities 2013: Conference Abstracts.

Gruszczyński, Włodzimierz, Dorota Adamiec and Maciej Ogrodniczuk. 2013. Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) – prezentacja projektu badawczego. *Polonica* XXXIII: 309–316.

Klemensiewicz, Zenon. 2002. *Historia języka polskiego*, Warszawa: Wydawnictwo Naukowe PWN.

Kurcz, Ida, Andrzej Lewicki, Jadwiga Sambor, Krzysztof Szafran and Jerzy Woronczak. 1990. *Słownik frekwencyjny polszczyzny współczesnej*. T. 1–2. Kraków: Polska Akademia Nauk.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski and Barbara Lewandowska-Tomaszczyk (eds.). 2012. *Narodowy Korpus Języka Polskiego*. Warszawa: Wydawnictwo Naukowe PWN.

R-manual 2015. R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/

Saloni, Zygmunt, Marcin Woliński, Robert Wołosz, Włodzimierz Gruszczyński and Danuta Skowrońska. 2015. *Słownik gramatyczny języka polskiego*. http://sgjp.pl/.

Woliński, Marcin. 2014. Morfeusz reloaded. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis (eds.). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1106–1111. Reykjavik: European Language Resources Association (ELRA).

Susan Conrad

Beyond Grammar Description: Applying Corpus Analysis to Disciplinary Education

Abstract Corpus-based studies of grammar have greatly increased our understanding of language use. As a field, however, corpus linguistics has been less successful in moving beyond description to substantive impacts. Many corpus studies claim important implications for education, but outside of second language teaching and translation, the results are rarely applied. In contrast, this paper describes a project designed to advance engineering education in the United States. The project has conducted several kinds of corpus-based grammar analyses of student and practitioner writing, and then applied the findings to materials that improve the preparation of students to write as professional engineers. Additional corpus analyses are used to analyze the impact of the materials on student writing. This paper traces the process used in the project and discusses its successes and challenges, encouraging other corpus linguists to apply their skills to diverse disciplines.

Keywords Corpus-based research applications, English corpus linguistics, engineering writing, corpus-based grammar teaching

1 Introduction

In recent decades, corpus-based analyses have contributed greatly to our understanding of English. Reference grammars produced since the late 1990s have differed greatly from traditional grammars that focused on accurate structure. For example, Biber, Johansson, Leech, Conrad and Finegan (1999) present over 300 analyses of variation in grammatical features' use, and McCarthy and Carter (2006) have chapters addressing spoken language and grammar, and utterances and discourse. A new generation of English as a second language (ESL) grammar textbooks also includes information about frequencies of features, patterns of lexis and grammar, and common learner errors. Most notably, Cambridge

University Press uses its Cambridge English Corpus seal on back covers of textbooks such as the recent *Grammar and Beyond* series (e.g., Reppen 2012), assuring readers that "you can be fully confident the language taught is useful, natural and fully up-to-date." Other publishers also offer corpus-based textbooks, such as Pearson Education's *Real Grammar* (Conrad & Biber 2009), which identifies itself as "a corpus-based grammar of English" that supplements traditional textbook information.

Other language-related fields have also been influenced by corpus linguistics work. In translation, for example, corpus-based studies have made it possible for the field to move from comparing single originals and their translations to examining – among other things – language patterns in translations more generally and translation-related shifts that occur regardless of the languages involved (see review in Bernardini 2015). Corpus techniques have been used in concrete applications in translation, not only advancing machine translation (e.g., Koehn 2005) but also providing a lexical and syntactic perspective for evaluating the quality of translations (Freire 2009).

Unfortunately, however, within education, corpus-based work has had little influence beyond language-centered fields such as translation and second language teaching. This is particularly surprising since there is ample evidence that almost all students – even native speakers – are challenged by the use of language as they enter a new discipline (see review in Wingate 2015). The findings of corpus-based analyses seem likely to be helpful for training in many disciplines, but impacts have been limited. Some corpus analyses that have included many disciplines are designed to be descriptive, not to have a direct application (e.g., Biber 2006). Other disciplinary work does have the potential for a direct application. For example, with a combination of corpus-based and experimental techniques in a study of German court decisions, Hansen, Dirksen, Küchler, Kunz, and Neumann (2006) found that reading comprehension was enhanced when the decisions were rephrased with simpler syntactic structures. They suggest their findings be used to teach law students. Few such implications become applications, however.

In this chapter, I urge corpus linguists to strive to have more impact – that is, to move beyond descriptive work into its application. I provide an example of a project that has used corpus analysis to examine an educational problem in the United States, to make teaching materials to address the problem, and to assess the effectiveness of the materials. The example demonstrates that, collaborating with disciplinary experts, corpus linguists can clarify and address student needs with great success.

In the next section I introduce the project, which focuses on civil engineering. I then present three corpus-based grammar analyses, illustrating different kinds of analyses that are useful in the project. Next, I exemplify how the analysis

results are applied in the development of teaching materials and briefly describe the additional corpus analyses that assess the outcomes from the new materials. The final section reflects on the project, highlighting characteristics that have made it successful and that are still challenging.

2 Civil Engineering and Corpus Linguistics

Most people come in contact with civil engineering every day through use of infrastructure such as roads, bridges, tunnels, water systems, buildings, and retaining walls. However, with the exception of engineers themselves, few people realize the important role communication plays in civil engineering. Studies within the industry have found that communication is the single most important factor in the success of infrastructure projects (Thomas, Tucker, & Kelly 1998) and poor communication has contributed to costly legal battles, structural failures, injuries, and deaths (Banset & Parsons 1989, Parfitt 2008, Parfitt & Parfitt 2007). Since large infrastructure projects are expensive and paid out of public tax funds, effective communication by engineers is also a financial concern for society. From a business perspective, too, writing is important; most firms' only product is written documents, and easy-to-understand writing is critical to clients' satisfaction and timely work.

There is a clear need, then, for civil engineering students to develop strong writing skills. In fact, this need has been discussed for decades, but employers and new graduates of engineering programs continue to express dissatisfaction with the preparation they receive (Berthouex 1996; Sageev & Romanowski 2001; Donnell, Aller, Alley & Kedrowicz 2011). The only studies of writing in engineering practice use surveys, small case studies, and anecdotal text evidence, and they rarely mention civil engineering (e.g., see Tenopir & King 2004, Winsor 2003, Sales 2006). Numerous textbooks for technical writing exist, but they have no empirical basis, and some studies have found they neglect the needs of engineering students (Wolfe 2009).

When I learned about the need to improve writing instruction within civil engineering, I immediately saw the usefulness of corpus linguistics to address this problem. With funding from the U.S. National Science Foundation and collaborators at three universities and in the local engineering community, I undertook a corpus-based project to investigate the gap between practitioner and student writing, clarify student needs, and develop materials to address the needs.

2.1 The Civil Engineering Writing Project

Figure 1 provides a schematic of the overall process in the Civil Engineering Writing Project.

The first phase, begun in 2009, compiled a corpus of 400 student papers from four universities and 400 workplace documents from 50 firms and agencies, covering ten registers (e.g. e-mails, technical memoranda, reports, plan sheet notes; see further Conrad, Pfeiffer & Szymoniak 2012). We then analyzed the corpus to investigate differences between student and practitioner writing. With the input of engineering practitioners in industry, we identified the most serious student writing weaknesses. In phase 2 of the project, currently underway, we develop teaching materials that address those writing weaknesses. In the intervention step, the materials are used in existing civil engineering courses. Students' papers from these courses – the post-intervention papers – are then analyzed and comparisons made with the pre-intervention papers, to assess the impact of the materials.

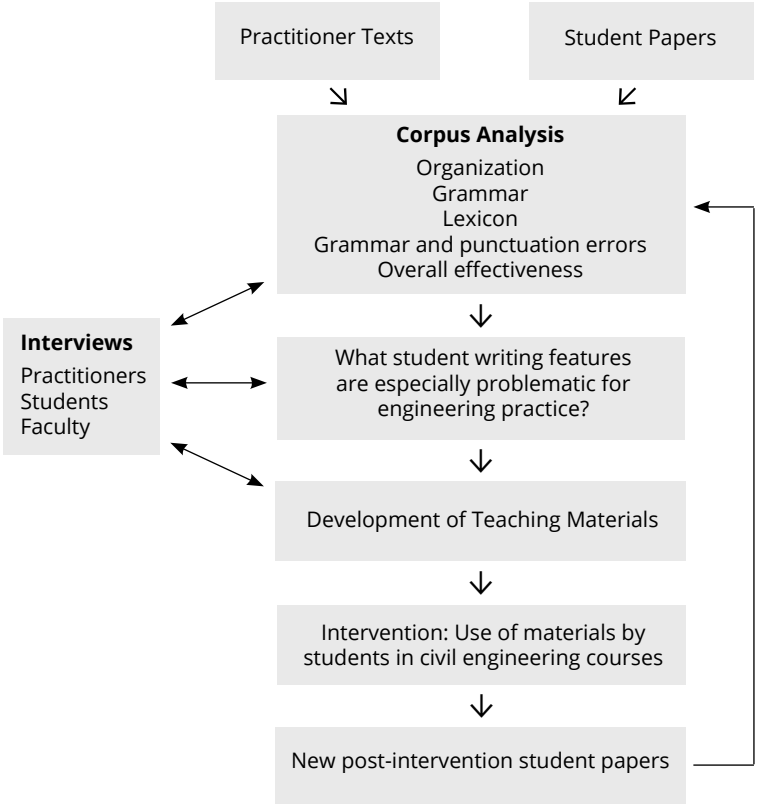


Figure 1: Overview of the Civil Engineering Writing Project process.

Three characteristics of the project might be surprising to readers more familiar with descriptive projects rather than teaching interventions. First, although the corpus has grown to over 1500 texts, the analyses typically focus on small subcorpora. The situational characteristics of many registers differ greatly (e.g., the content, communicative purpose, and audience of a student lab report are very different from a practitioner design report) and an overall description of the linguistic variation – though interesting to linguists – is not especially helpful for designing teaching materials.

A second notable characteristic of the project is the interplay of the corpus analysis with interview data. Corpus projects often consult disciplinary experts for corpus design issues or to understand disciplinary conventions, but this project relies even more heavily on input from practitioners and students. An especially useful step has been sharing the results of corpus analysis with interviewees. Student reactions help us to understand the "why" behind their writing choices, something no corpus analysis can reveal. Practitioner explanations allow us to understand which student writing problems are the most important to address and which changes in student writing are most effective. Practitioners also contribute to the teaching materials, commenting on drafts and checking that all information – even if it is simplified for a beginning-level course – is consistent with engineering practice. The examples in the next sections share some specific contributions from interviews, based on interviews with 22 students and 16 practitioners. (Faculty are also interviewed but are not the focus of this paper.)

The third characteristic concerns the diversity of the universities who participate in the project. Compiling a corpus from multiple universities is more time-consuming than focusing on just one, but for this project it was crucial for identifying weaknesses shared by different student populations and investigating the impact of the materials with diverse groups. The project is based at Portland State University in the northwestern U.S. and includes three other universities: the California State Polytechnic University at Pomona, Howard University in Washington, D.C., and Lawrence Technological University in the Midwest. All offer an accredited Bachelor's degree in civil engineering and seek to train students to become effective practitioners, but they differ in size, geographic region, entrance requirements, and typical student academic and ethnic backgrounds.

3 Grammar Analyses

This section summarizes three of the grammar-related analyses from the first phase of the project, which revealed differences in student and practitioner writing and also challenged many claims about engineering writing. I highlight just a few of the most important aspects of the analyses; further details about the

methods and results can be found in other publications about the project, especially Conrad (2015, 2017, and 2018).

3.1 Passives and Impersonal Style

It is widely claimed that engineers overuse passive voice and make texts too impersonal. For example, Gwiasda berates the high frequency of passive voice in student writing as "the perfect vehicle for documents that record material of no intended consequence to anyone at all" (Gwiasda 1984: 150). Sales (2006: 18) describes practicing engineers as "consciously avoiding any use of the personal pronouns" in order to be more objective. There is no systematic evidence to support these claims, but previous corpus-based investigations of academic prose (e.g., Biber 1988) have found engineering to use a higher frequency of passives than most academic texts.

For an analysis of passives and impersonal style features in the civil engineering texts, I used a sub-corpus chosen so that practitioner and student writing was as similar as possible and represented a typical workplace writing task – reports written to clients, addressing real situations (Table 1). This is a task typically given to students in their fourth (final) year of the degree. For a comparison with professional academic texts, I also included 50 research articles.

Table 1: Texts used in the passive voice and impersonal style analysis.

Category	Number of texts	Sources	Words
Practitioner Reports	60	10 firms	201,700
Student Reports (for clients)	60	9 courses	207,700
Journal Research Articles	50	10 journals	270,900

The analysis used a technique well established in corpus linguistics – Multidimensional (MD) analysis, as introduced by Biber (1988). MD analysis uses a factor analysis to calculate the co-occurrence patterns of linguistic features in texts. Groups of features that tend to occur together in texts are identified statistically; no a priori assumptions are made about which features should be grouped together. The factors are interpreted in terms of their communicative functions as dimensions of register variation. In the study of 23 registers of spoken and written English conducted by Biber (1988), one factor had four kinds of passive structures – agentless passives, passives with *by* prepositional phrases, past participial clauses, and past participial noun postmodifiers (Table 2). In addition, two kinds of connecting words loaded onto the same factor: linking adverbials and multi-functional subordinators. This dimension was characterized as

Table 2: Features on the Impersonal Style dimension.

Language Feature	Example	Factor loading
linking adverbials	*therefore, however, in conclusion*	.48
passive verbs, agentless	The bridge *was built* in 1923.	.43
past participial clauses	*Designed by a local engineer*, the bridge won an international award.	.42
passive verbs with *by* phrases	The bridge *was designed* by a local engineer.	.41
past participial noun postmodifiers	The recommendations *included in this report* cover …	.40
adverbial subordinators with multiple functions	*since, while, whereas, such that*	.39

Impersonal Style, reflecting the high frequency of passives and lack of human agents. The connectors were found to overtly structure the logical relationships in the often dense, technical texts. I applied this dimension for the analysis of the engineering texts.

I used the standard procedures for the MD analysis as outlined in Conrad and Biber (2001). I grammatically "tagged" the files with the Biber tagger and checked and corrected features with another program. Grammatical features in the engineering registers were counted and standardized to the findings of Biber's (1988) analysis so that comparisons could be made with a range of English discourse. In Figure 2, which displays the results of the analysis, 0 represents the mean for the 23 registers in Biber's analysis, and each positive or negative unit represents a standard deviation.

As Figure 2 shows, the results of the analysis are generally consistent with claims that engineering writing is highly impersonal; relative to a wide range of English discourse, the three registers of engineering all have a markedly high mean score on the Impersonal Style dimension. Their use of impersonal features is, for example, far higher than conversation, fiction, and popular nonfiction (magazines and books for a non-specialist audience). However, when the engineering registers are compared among themselves, the differences are important. An analysis of variance found a statistically significant difference among the three engineering registers ($F(2, 167) = 19.89$, $p < .0001$, $\eta^2 = .19$), with the student papers and journal articles using more impersonal style features than the practitioner papers. Post-hoc Scheffe pairwise comparisons found a statistically significant difference between the practitioner reports and student reports, and between the practitioner reports and journal articles, but not between the student reports and journal articles. In other words, in the frequency of impersonal style features, the student reports resemble academic journal articles more than the practitioner reports they are meant to imitate.

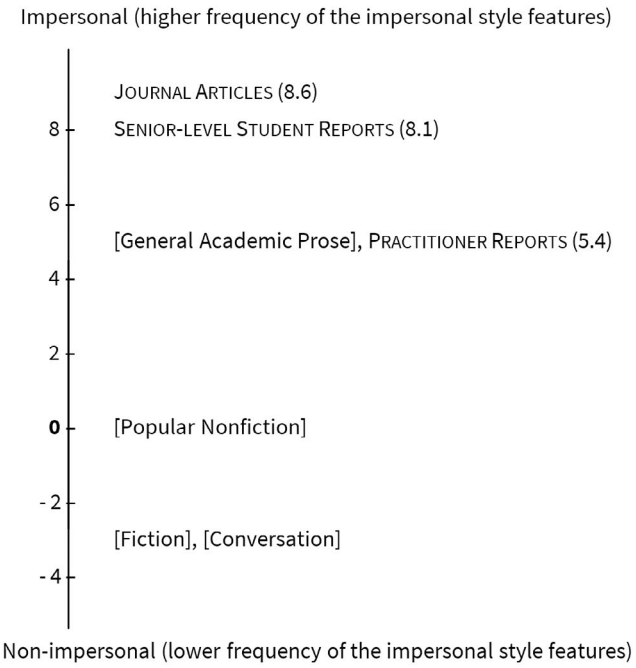


Figure 2: Mean scores for three civil engineering registers on the Impersonal Style dimension. *Note:* General academic prose, popular nonfiction, fiction, and conversation are from Biber (1988) for comparison.

Several of the important characteristics from the Impersonal Style analysis are exemplified in this excerpt from a practitioner report:

(1) On August 15 and 19, 2003, we drilled five exploratory borings with a portable drill rig using solid stem auger techniques. These borings were drilled to provide data for retaining wall and signal pole foundation design. The boreholes were drilled to depths ranging from ±2 to 6 m.

Surprisingly, the paragraph begins with a human agent and active voice (*we drilled*). Although not as common as passives, these structures appeared regularly in practitioner texts with a variety of verbs (*we observed..., the subject team conducted..., ABC Engineering recommends..., we anticipate...*). In interviews, practitioners commented that occasional overt statements of responsibility were important; they not only made it "easy for readers to read fast" but they were important to "manage liability in a field where you are hired for subjective judgments." Contrary to the claims in the literature about engineers seeking to sound objective, these practitioners emphasized making subjective judgments based on

observed data. They discussed the need to be explicit about responsibility for observations and judgments. They especially emphasized being explicit about recommendations because recommendations from a licensed engineer have a legal status; they must be followed unless they are changed by another licensed engineer.

The second and third sentences in example 1 use passives. They illustrate three functions that commonly occur with passive voice. First, they allow objects, processes, or concepts to be the grammatical subject and thus a consistent topic of discourse (here: *these borings, the boreholes*). Second, the passive constructions conform to the principles of information structure and end weight (Biber et al. 1999). That is, the subject noun phrases in the passives refer back to the topic established in the previous sentence (*borings*), and the information after the verb (*to provide data for..., to depths ranging...*) is new information that is longer than the subject noun phrase. Only the first of these three functions is typically mentioned in technical writing materials even though conforming to typical information structure and end weight can be crucial for making technical information easy to read.

An additional important characteristic that accounted for fewer passives in practitioner writing was the more frequent use of inanimate subjects with active voice verbs. Objects, processes, and documents often do things in these engineering texts – for example, *this document reports the analysis...* and *our analysis assumes a factor of safety of....*

The journal articles and student papers used passives more consistently. Passives were regularly used for the kind of actions practitioners expressed in active voice, such as recommendations and observations:

(2) a. It is recommended that these new equations and charts should be included in the revision of the AASHTO Bike Guideline. (journal article)
b. Due to the design of the intersection, initially it was thought that cyclists would merge to the right lane and be forced to compete with merging freeway traffic, but it was observed that most cyclists merged safely into the left car lane well before reaching the intersection. (student report)

Since recommendations in journal articles do not entail any legal meaning, the lack of explicit responsibility and use of the hedge *should be* do not have a critical impact, as they might in a practitioner report. The writing in (2b), however, is meant to imitate a practitioner report. Instead, its passives leave the reader wondering who is responsible for this work: what mysterious group was hypothesizing about how cyclists will merge to the right lane? And was it that group or another who observed the cyclists merging safely? The difference from

practitioner reports is striking, but in interviews, most students said they had learned that technical writing should not use personal pronouns or refer to people. They commented on "...the technical writing thing of don't use *I* or *we* or *us*" and stated "You need to use objective language." Some writers clearly thought the absence of human agents automatically created objective meaning; they used expressions such as *it was believed...* or *it was felt...*, but – even in passive voice – beliefs and feelings are not appropriate evidence for engineering.

When students were shown examples like (2b) in interviews, many also commented that they used such sentences because they were long or looked "fancy." This desire to look fancy also contributed to a high frequency of linking adverbials and subordinators in texts. Unfortunately, the fancy sentences were also often ineffective; in (3), for instance, the important conclusion – the recommendation to use bike lanes and bioswales – is minimized by being in a subordinate clause:

(3) ... *Moreover*, SW Elm is fully paved with standard asphalt (highly impermeable) and relies fully on gutters to carry off rainwater. *Thus*, water overflow can occur on the site during heavy rain seasons, *while* having permeable pavements and bioswales could solve this issue.

The analysis of the impersonal style features added to our understanding of student and practitioner writing in notable ways. It countered the image of all engineering writing being like academic writing; in fact, workplace writing incorporates more human agency because explicit responsibility and unambiguous content is valued. It provided systematic evidence for claims that passives are often useful in writing that focuses on objects, but it also highlighted passives' usefulness for conforming to typical end weight and information structure. It also revealed the student's weakness for "fancy" sentences, which is taken up in the next analysis.

3.2 Sentence Structure

Another widespread belief about engineering writing is that sentences are needlessly long and complicated. An online website for career and education information for a professional engineering society, for example, quotes a technical writing consultant with 25 years of experience: "I have met very few engineers who are comfortable with using simple language, organizing documents for the readers' benefit, keeping sentences and paragraphs short, and getting to the point" (Crawford 2012: 2).

One approach for investigating sentence complexity in corpus-based studies is to use automatic counts of complexity features, but in pilot work we found

that some student texts had such numerous sentence structure and punctuation errors, it was difficult to automatically identify clause structure. For this example, then, I illustrate a different kind of analytical technique that is useful in the project – coding a sample by hand.

For the sentence structure analysis, we sampled sentences in the texts in Table 3. Originally interested in development as students progressed in their major, we included third-year student lab reports, the most common type of third-year writing students do. For fourth-year students and practitioners, we included reports and technical memoranda – two registers that are common in the workplace and final-year courses. It turned out that preliminary analyses found no difference in the two student groups, so they were combined in the analysis reported here.

Table 3: Texts used in the sentence structure analysis.

Category	Number of texts	Sources
Practitioner reports and technical memoranda	86	10 firms + 1 public agency
Student reports and technical memoranda (senior level)	78	9 courses
Student laboratory reports (junior level)	122	4 courses

For the analysis, I made a simple distinction between sentences that were "complicated" or non-complicated, defining complicated as having dependent or embedded clauses. The more detailed categories typical of linguistic studies, such as finite versus nonfinite dependent clauses or postnominal versus adverbial clauses, were more specific than needed for the general comparison of sentence complexity we sought and too detailed for the engineers to understand quickly.

I followed a standard procedure of multiple samples, often used in corpus-based studies that require hand-coding of data (Biber, Conrad & Reppen 1998: 91–93). Specifically, for each of the writer groups, I analyzed three random samples of 100 sentences. The proportions of complicated sentences was within 5 % for each sample, so I took them as representative of the group. The complete sample was thus 600 sentences.

A chi-square test found a statistically significant difference between the frequency of complicated sentences in the practitioner and student writing (χ^2 = 51.3, df = 1, p < .001, φ= .293) with the students using more complicated sentences. Over half of the student sentences had complex or embedded structures, while only about a quarter of the practitioner sentences did (Figure 3).

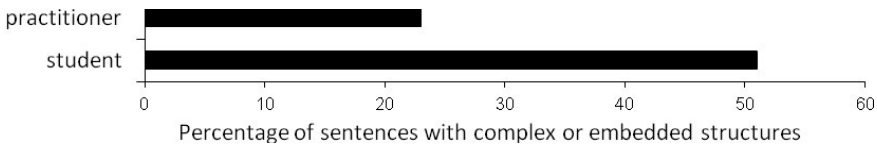


Figure 3: Use of complicated sentence structure by students and practitioners.

Practitioner writing had more sentences expressing a single idea, as in the following examples:

(4) a. The rainfall depth was obtained from the City of Granson, County of Wilson. For the 25-year storm event, 24-hr rainfall depth is 4.0 inches for the site.
b. The lower portion of the embankment, below ±El. 475 to 480 and near Harmony Creek, is graded at approximately 1½(h):1(v).

Sentences like (4b) look long to students and might contribute to student beliefs about "fancy" sentences. However, linguists can easily see that the length comes from phrasal complexity, especially long noun phrases and prepositional phrases that make information very precise (see further discussion in Conrad 2015: 325–6). The clause structure remains simple. Commenting on the frequency of simple sentence structures, practitioners again noted the need to make information as easy as possible for clients to follow. They commented, for example, "Clients want to be able to read fast or skim," and "Simple sentences are more concise. And they are less likely to be ambiguous or be misinterpreted."

Student sentences, on the other hand, tended to have more complexity on the clausal level, as illustrated in this sentence from a transportation report, which has multiple clausal constituents and one subordinate clause embedded within another subordinate clause:

(5) [This particular modeling detail does not seem [to greatly affect the output of the simulation] [because [although it appears unrealistic], it does not affect the flow of traffic greatly and only seems [to occur on occasion]]].

Such student sentences are, at best, hard to follow. Sometimes they even became so complicated that their literal meaning was inaccurate. In interviews, however, students expressed no concern for making texts easy to read and unambiguous. Instead, when students were asked to comment on complicated sentences, typical explanations for choosing them were:

"It looks better if it's longer. I think it's that simple."
"Make it fancy."

"I kind of felt like I had to sound professional and smart. I mean, you want to sound really knowledgeable about things, and it seems like the easiest way to do that is to be wordy."

Overall, this analysis was useful because it provided systematic evidence that it is students – not practitioners – who write with complicated sentence structures. The interviews made clear that practitioners valued the simpler clause structure for their ease of reading and the complex phrases for the specificity of information. The analysis provided evidence that students' writing and their beliefs about writing were the opposite of practitioners'.

3.3 Errors in Grammar and Punctuation

Initially, I did not plan to include error analysis in the project because grammatical choices and their impacts, not basic accuracy, seemed most important for writing. However, it was soon obvious that errors had a large impact on students' writing effectiveness. Furthermore, several civil engineering faculty firmly believed that it was only ESL papers that had a high frequency of errors when I suspected errors were more widespread. I therefore added an error analysis to the project.

The analysis investigated the extent to which writers conformed to standard written English grammar and punctuation. It followed procedures for hand-coding errors as in traditional learner corpus studies. Because the coding of errors is time-consuming, the analysis covered a subset of the papers in Table 3 (above), using 45 texts each from the practitioners, senior-level students, and junior-level students. The senior-level and junior-level papers were counted separately since the frequency of errors varied greatly.

Errors were categorized into five major categories (Table 4) by trained research assistants. The errors typical of ESL students provided a rough means of assessing whether ESL-type errors dominated the analysis. Native speakers of English also make these kinds of errors, but they tend to be more common in ESL texts.

Table 4: Error categories in the error analysis.

Error Category	Description
1. Verb errors	Tense, aspect, formation of infinitives and other verb forms, any verb errors other than S-V agreement
2. Sentence structure	Any structure errors that make sentence ungrammatical in English, includes relative clause or participle clause errors
3. Punctuation	Commas, semi-colons, sentence-final punctuation, and other punctuation
4. Spelling and typos	Errors related to spelling or typing
5. Articles, prepositions and other errors typical of ESL learners	Errors with articles, prepositions, plurals, subject-verb agreement and pronoun-antecedent agreement

Errors in each category and total errors were counted per text and normed per 1,000 words. Figure 4 displays the median error frequencies across the groups: just over 2 for practitioners, about 13 for senior-level papers, and almost 16 for junior-level papers. On a double-spaced, printed page, these frequencies mean about one error on every other page for practitioner documents, about three per page for senior-level papers, and about five per page for junior-level lab reports. A Kruskal-Wallis one-way analysis of variance test found a significant difference in the three groups' error rates overall ($H(2) = 60.855$, $p < .001$). There was a statistically significant difference between the practitioner writing and senior-level writing ($p < .001$, $r = 0.67$) and between the practitioner writing and junior-level writing ($p < .001$, $r = 0.75$), but not between the senior-level and junior-level writing.

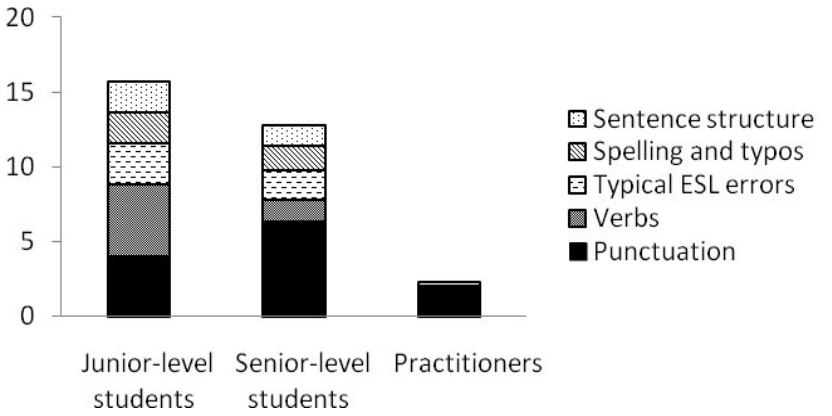


Figure 4: Median error rates in student and practitioner writing.

Although a few student papers were almost error-free, the median rates show that many student papers had enough errors to be distracting and damaging to the writer's credibility. The errors were also more widespread than ESL students would account for, especially since the senior-level papers were written in groups and interviewees commonly reported that native English speakers edited ESL writers' contributions.

The student and practitioner texts also differed in the types of errors they included and their impacts on comprehensibility. In the practitioner documents, punctuation accounted for the vast majority of errors, as Figure 4 shows. The majority of these errors involved isolated comma errors that did not interfere with meaning. Student errors, on the other hand, covered all categories. Some errors were just odd, such as unusual punctuation choices (example (6a)), perhaps related to the desire to "make it fancy." Some errors made sentences literally nonsensical, such as the dangling modifier in example (6b). The most serious usually involved sentence structure errors and made the main idea difficult to discern, as (6c) exemplifies.

(6) a. The map displays the geologic conditions; with the basalt layers in darker colors.
b. As a civil engineer, the strength of concrete is highly affected by the curing time.
c. But the brittleness of each coupon varied with coupon #3 having little necking and being the most brittle of the three coupons, coupon #13 had more necking than #3 but less than #7 and thus concluding it had moderate ductility of the three coupons.

When discussing errors, practitioners' most common comment had to do with engineering being a detail-oriented profession. They were concerned about errors inadvertently changing meaning and also making the firm look unprofessional. One interviewee summed up a credibility problem for the writer: "Errors convey carelessness. Who wants a careless engineer?" Some mentioned that they were shocked by the level of errors in some job applications they received and that those applications went straight into the trash.

All the students said they proofread their papers at least once, but many reported spending little time because they perceived errors to have little influence on their grade. This perception was consistent with a review of lab reports that received grades of 90% or above; they included papers with some of the lowest and highest error rates. Many students also reported that, even when they did proofread thoroughly, they had little confidence in their ability to recognize and correct errors.

This analysis provided evidence to counter the faculty impression that errors are a problem only for ESL students. They are a serious problem for many students. They also constitute a serious matter for the practice of engineering. Errors can undermine the credibility of a new graduate applying for a job, a practicing engineer, or the professional reputation of a firm.

4 Applying the Corpus Research to Improve Teaching

The results of the analyses were used to develop the new teaching materials. These materials are free-standing units that cover genre expectations, grammatical and lexical choices, and grammar and mechanics errors. This section uses examples of the materials related to the grammar analyses described above. More details can be found at the Civil Engineering Writing Project website, www.cewriting.org, and in Conrad, Kitch, Smith, Lamb & Pfeiffer (2016).

4.1 Features of the New Teaching Materials

Each unit is drafted by applied linguistics and engineering faculty and is then reviewed by at least two practitioners, who check that advice is consistent with workplace practice. Here I highlight four features that set the materials apart from typical technical writing instruction, made possible by the combination of the corpus analysis and interview data.

First, the units provide information about the patterns of language features that differ between student and practitioner writing and, with practitioner quotes, tell why the language features matter within civil engineering practice. The opening of a sentence structure unit illustrates these features (see appendix). Students see a figure comparing the percentage of simple sentences in student and practitioner reports. The findings are described for the students, and the target for revising is explicit (use more sentences that express one idea). The importance of simple sentences for engineering practice is reemphasized by comments from practitioners.

Each unit also contains numerous examples of practitioner writing. For many students, this is a first experience seeing sentences from practitioner documents. We choose examples that illustrate the most important corpus findings. We also provide explanations that use simple terms to direct students' attention to linguistic features. Figure 5 provides an example from the unit about simple sentence structure.

Effective Simple Sentence Structure	
Examples	Explanation
1. **The existing bridge is** a 9-span timber trestle bridge with a concrete deck. **It is** 217 feet long and 30 feet wide. **The posted speed is** 25 mph. *(Report)*	Each sentence has one main idea. It has a subject (in purple) and a verb phrase (in red). The verb is close to its subject.

Figure 5: Opening of a section exemplifying and explaining practitioner writing.

Many units also contain "Myth buster" boxes. These boxes present information that directly counters the misconceptions that students expressed in interviews and that underlie ineffective writing choices. For example, the unit about passive voice counters the idea that passive voice automatically expresses objectivity (Figure 6). It addresses the fact that engineering requires judgment and ties it to the use of human agents with active voice. It goes on to urge students to strive for accurate meaning in verbs, rather than relying on passives such as "it was felt that..." since "feeling" is not adequate evidence in any voice.

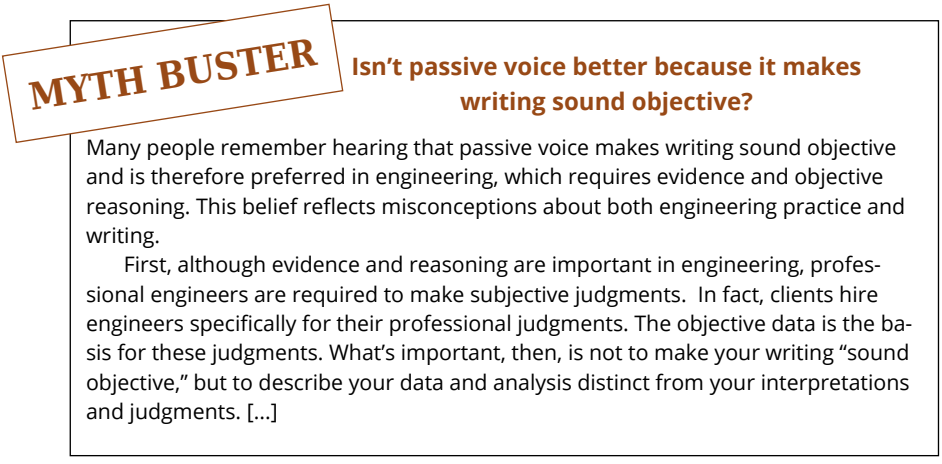


Figure 6: Example of a "myth buster" box from the passive voice unit.

The units also cover specific revision techniques and provide practice activities for them. This kind of practice is not unusual in writing materials, but using the corpus allows us to include real student sentences, and give students realistic revising practice that addresses common problems. The unit on passive voice, for example, includes tips on using inanimate subjects with active verbs (Figure 7). The units that address grammar and mechanics address the most common

errors, some of which – like the overuse of semi-colons – would not have been recognized without the corpus analysis.

Technique 4: Use an inanimate subject + active voice verb.	
Original Sentence Needing Revision	Revision
1. [*Note: preceding paragraph describes the basis for the liquefaction analysis]* A potential for liquefaction in the loose sand between 15 and 30 feet was indicated. *(Report)*	1. The results of the analysis indicate a potential for liquefaction in the loose sand between 15 and 30 feet.
Explanation. The original of example 1 has a long subject before the verb. The revision uses a shorter, inanimate subject + active verb (*results indicate*) for easier reading. The revision also now follows expected information structure in two ways: it explicitly moves from data analysis to the engineers' interpretation of it (see Unit 4, Part 1) and it follows known-new information sequencing (see Unit 4, Part 2).	

Figure 7: Example revision technique for reducing overuse of passive voice.

4.2 Assessing the Effectiveness of the Materials

After the materials are used in courses in civil engineering departments, students write papers that are compared to pre-intervention student papers. Currently, we have results from four universities, three levels (first-, third- and fourth-year courses), and 16 different courses. The materials have been implemented in a variety of conditions. Class size has ranged from 12 to 80 students. The amount of class time versus homework time for the materials has varied from a writing workshop day in class to no class time at all. Some courses had writing teaching assistants; most did not. Although this variability can make assessment more challenging, we want the materials to be piloted in realistic conditions.

The same techniques used for analyzing differences in practitioner and student writing are used to analyze the change in student papers. This includes the techniques described above, plus a separate analysis of passive main verb effectiveness, word choices, and genre organization (further information can be found in Conrad, Kitch, Pfeiffer, Smith, & Tocco, 2015). In addition, the assessment includes a holistic evaluation of effectiveness by a practitioner since changes in linguistic forms do not always amount to an improvement in overall effectiveness. The results are summarized in Table 5, with the grammar features described in this paper in the top half of the table, and other features in the bottom half. As the summary in the table shows, the results have been consistently positive.

Table 5: Summary of post-intervention results (16 courses).

Language feature	Change in student writing
Passive Voice	Statistically significant reduction in frequency of passive voice Active voice used appropriately for responsibility
Sentence Structure	Statistically significant reduction in complicated sentences No complicated sentences with inaccurate meaning
Grammar and Punctuation Errors	Statistically significant decrease in targeted errors Decrease in errors that interfere with meaning
Word Choices	Statistically significant reduction in vague or inaccurate words
Genre Analysis (organization)	Statistically significant increase in effectiveness of content sequencing, inclusion of expected content, and decrease in extraneous content
Evaluation by Practitioner	Statistically significant increase in overall effectiveness rating

We also ask students for their reflections and suggestions after they use the materials. Their reflections show that the materials can impact attitudes and beliefs that underlie some of the ineffective features of student writing. Typical comments have included the following:

"The information that made the biggest impression on me was that engineering writing is different from literature writing and can cost me a job."
"The thing that impressed me most today was how poor my grammer [sic] and editing skills are."
"I think the biggest challenge for me in writing for CE will be to ignore the temptation to sound fancy and smart."

The only consistent suggestion we have received is to include more examples even though the units are already longer than we planned for easy incorporation into courses.

Of course, the positive results of the assessment do not mean every post-intervention student paper is strong. In fact, it occasionally appears that a student did not look at an assigned unit at all. Certain individuals, for example, never stop overusing complex sentences, and we hope to investigate this individual variation more in the future.

5 Conclusion

The evidence from the Civil Engineering Writing project suggests that corpus-based grammar description can indeed be applied to have positive impacts in disciplinary education. To conclude, I reflect on some of the most important factors for the success of the project and others that continue to be our biggest challenges.

One characteristic that contributes to the success of the project is the highly specialized nature of the corpus. Even if the corpus focused on all engineering rather than only civil engineering, it would be impossible to identify student weaknesses as specifically because work contexts could vary so greatly. It is even more important that we were able to compile a corpus to represent the kind of workplace writing students hope to do after graduation, not just academic writing. Compiling a corpus of workplace texts is easier in civil engineering than many fields because the documentation of any publicly funded project is open to the public; in many other fields, issues of confidentiality would likely make corpus compilation more difficult.

Civil engineering is also well suited to a corpus-based project because the field is data-oriented. Engineers expect to see data analysis, especially quantitative data, as a basis for decision-making. Even if they do not understand all the linguistic details of an analysis, they generally appreciate the quantitative evidence in conjunction with explanations of language functions. Other fields in science, technology, engineering and mathematics are likely to be equally appreciative partners in a corpus project, but some other fields might consider the quantitative analysis less valuable.

Success has also depended on having access to helpful disciplinary experts. Numerous practitioners have been generous with their time, both in teaching me about civil engineering generally and in answering numerous writing- and language-related questions. They are aware of how important writing skills are in their profession, and many struggled in their own first attempts to write in industry. Without their input, we simply could not target workplace writing skills as we have.

Project success is also dependent on civil engineering faculty, who help develop the materials and try them in their courses. Many faculty have contributed, but this continues to be one of the most challenging aspects of the project. Most faculty have no training in teaching writing, nor do they have any metalanguage for explaining language choices. Even those who are enthusiastic about using the materials in courses admit it takes some time to be comfortable with them and to feel prepared to answer the kinds of questions students typically ask. Many faculty also find it challenging to add anything more to their already full syllabi. A number of faculty are resistant to using the materials at all. A

shortcoming of the project is that I did not plan faculty training seminars, which would likely increase enthusiasm for using the materials.

Finally, another continuing challenge in the project concerns teaching linguistic phenomena to an audience that generally has little language training and little metalanguage for referring to language. In materials, it is often difficult to be accurate about linguistic phenomena, but also easy enough for the audience to understand. Even referring to sentence structure is difficult because terms like phrases, clauses, and subordination are not known. Effective descriptions often require multiple rounds of drafts, feedback, and revisions. I also find it a satisfying challenge, however, because people untrained in linguistics learn to recognize how to manipulate language in more effective ways and even how to explain effective choices to each other.

All of these factors – and others – make an applied, corpus-based project challenging. Nonetheless, I have found any aggravations well worth seeing the improvements in student writing. Corpus-based descriptions provide a basis for work that other approaches cannot match. I urge other corpus grammarians to consider the wider audiences who might benefit from the applications of their work and to start working with them. Otherwise, though corpus linguistics will continue to be known within linguistics and language studies, it will not help to solve problems in other disciplines, where corpus analysis can make such a valuable contribution.

Acknowledgements

Partial support for this project is provided by the U.S. National Science Foundation (awards DUE-0837776 and DUE-1323259). I am also grateful to the civil engineering practitioners, students, and faculty who make the project possible.

Appendix – Example opening of a unit about sentence structure

Civil Engineering Writing Project – Language Unit 3
EFFECTIVE SENTENCES: SIMPLE SENTENCE STRUCTURES

What do you need to know about effective writing in civil engineering practice?

Experienced engineering practitioners use **simple sentence structure** in most of their writing. Simple sentence structure is effective because it conveys one main idea. Simple sentence structure makes comprehension easier for readers especially when sentences have complex, precise technical information.

Students use fewer simple sentences than practitioners do (Figure 1). In other words, students use complicated sentences more often. Students' sentence structure is more similar to academic journal articles than practitioner documents. In addition, students' complicated sentences often make content ambiguous or inaccurate. Revising sentence structure can therefore be an important step towards effective writing.

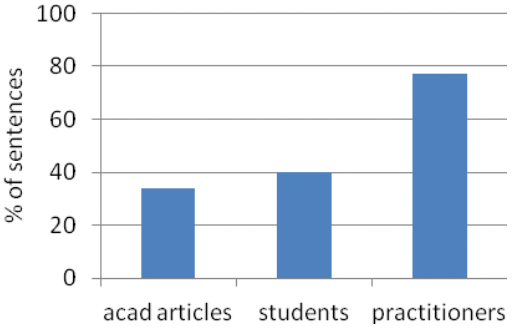


Figure 1: Percentage of sentences with **simple sentence structure** in student reports, practitioner reports, and academic journal articles

What experienced engineering practitioners say

"Clients want to be able to read fast or skim."

"Simpler sentences are more concise. And they are less likely to be ambiguous or be misinterpreted."

References

Banset, Elizabeth and Gerald Parsons. 1989. Communications failure in Hyatt Regency disaster. *Journal of Professional Issues in Engineering* 115: 273–288.

Bernardini, Silvia. 2015. Translation. In Douglas Biber and Randi Reppen (eds.), *The Cambridge handbook of English corpus linguistics*, 515–536. Cambridge: Cambridge University Press.

Berthouex, P.M. 1996. Honing the writing skills of engineers. *Journal of Professional Issues in Engineering Education and Practice* 122: 107–110.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2006. *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *The Longman grammar of spoken and written English*. Harlow, UK: Pearson Education.

Conrad, Susan. 2015. Register variation. In Douglas Biber and Randi Reppen (eds.). *The Cambridge handbook of English corpus linguistics*, 309-329. Cambridge: Cambridge University Press.

Conrad, Susan. 2017. A comparison of practitioner and student writing in civil engineering. *Journal of Engineering Education* 106: 191–217.

Conrad, Susan. 2018. The use of passives and impersonal style in civil engineering writing. *Journal of Business and Technical Communication* 32: 38–76.

Conrad, Susan and Douglas Biber (eds.). 2001. *Variation in English: Multi-dimensional studies*. Harlow, UK: Pearson Education.

Conrad, Susan and Douglas Biber. 2009. *Real grammar: A corpus-based approach to English*. White Plains, New York: Pearson Education.

Conrad, Susan, William Kitch, Timothy Pfeiffer, Tori Smith and John Tocco. 2015. Students writing for professional practice: Collaboration among faculty, practitioners, and writing specialists. *Proceedings of the 2015 American Society for Engineering Education Annual Conference and Exposition*. http://dx.doi.org/ doi:10.18260/p.24769.

Conrad, Susan, William Kitch, Tori Smith, Kenneth Lamb and Timothy Pfeiffer. 2016. Faculty-practitioner collaboration for improving civil engineering students' writing skills. *Proceedings of the 2016 American Society for Engineering Education Annual Conference and Exposition*. http://dx.doi.org/10.18260/p.26892.

Conrad, Susan, Timothy Pfeiffer and Thomas Szymoniak. 2012. Preparing students for writing in civil engineering practice. *Proceedings of the 2012 American Society for Engineering Education Annual Conference and Exposition*. https://peer.asee.org/21817.

Crawford, Mark. 2012. How engineers can improve technical writing. https://www.asme.org/career-education/articles/business-writing/how-engineers-can-improve-technical-writing.

Donnell, Jeffrey, Betsy Aller, Michael Alley and April Kedrowicz. 2011. Why industry says that engineering graduates have poor communication skills: What the literature says. *Proceedings of the 2011 American Society for Engineering Education Conference and Exposition*. https://peer.asee.org/18809

Freire, Evandro. 2009. A corpus-based approach within Juliane House's model for translation quality assessment. *The ESPecialist* 30: 193–211.

Gwiasda, Karl. 1984. Of classrooms and contexts: Teaching engineers to write wrong. *IEEE Transactions on Education* E-27: 148–150.

Hansen, Sandra, Ralph Dirksen, Martin Küchler, Kerstin Kunz and Stella Neumann. 2006. Comprehensible legal texts – utopia or a question of wording? On processing rephrased German court decisions. *Hermes – Journal of Language and Communication Studies* 36: 15–40.

Koehn, Philipp. 2005. EuroParl: A parallel corpus for statistical machine translation. *Proceedings of the Tenth Machine Translation Summit*, 79–86. http://mt-archive.info/MTS-2005-Koehn.pdf.

McCarthy, Michael and Ronald Carter. 2006. *Cambridge grammar of English*. Cambridge: Cambridge University Press.

Parfitt, M. Kevin. 2008. Building failures: Avoiding the mistakes of yesterday leads to the successes of today. *Journal of Architectural Engineering* 14: 31.

Parfitt, M. Kevin and Elizabeth Parfitt. 2007, January. Failures education: The key to better engineering design. *Structure Magazine*, 10–12. http://www.structuremag.org/wp-content/uploads/2014/09/p10-12C-EdIssues-FailuresEducation-Jan071.pdf.

Reppen, Randi. 2012. *Grammar and beyond 2*. New York: Cambridge University Press.

Sageev, Pneena and Carol Romanowski. 2001. A message from recent engineering graduates in the workplace: Results of a survey on technical communication skills. *Journal of Engineering Education* 90: 685–693.

Sales, Hazel. 2006. *Professional communication practices in engineering*. Basingstoke, UK: Palgrave Macmillan.

Tenopir, Carol and Donald King. 2004. *Communication patterns of engineers*. Hoboken, New Jersey: John Wiley/Institute of Electrical and Electronic Engineers.

Thomas, Stephen, Richard Tucker and William Kelly. 1998. Critical communication variables. *Journal of Construction Engineering and Management* 124: 58–66.

Wingate, Ursula. 2015. *Academic literacy and student diversity*. Bristol, UK: Multilingual Matters.

Winsor, Dorothy. 2003. *Writing power: Communication in an engineering center*. Albany, New York: State University of New York Press.

Wolfe, Joanna. 2009. How technical communication textbooks fail engineering students. *Technical Communication Quarterly* 18: 351-375.

II. Methodology and Application
Current Trends and Issues

Tassja Weber

Grammatik und Lernerkorpora: Eine korpusorientierte Untersuchung von Präpositionalphrasen im deutschen MERLIN-Korpus

Abstract This pilot study using the German learner corpus MERLIN aims to explore the impact of syntactic functions of prepositional phrases (PP) on the use of prepositions by learners of German as a foreign language. The paper focuses on complements containing specified prepositions licensed by verbs and adjectives, and adjuncts (as well as adjunct-like complements) containing unspecified prepositions. The frequent German prepositions *an* (*at*) and *auf* (*on*) were extracted from the learner corpus and the PPs annotated according to their syntactic functions. Results show that specified prepositions lacking semantic content seem to pose significantly greater problems to learners. Additionally, prepositions are omitted significantly more often in complement-PPs than in adjunct-PPs.

Keywords Grammatik, Lernerkorpora, Präpositionen, Präpositionalphrasen, Annotation, Deutsch als Fremdsprache

1 Einleitung

Lernende des Deutschen als Fremdsprache (DaF) zeigen Schwierigkeiten beim Gebrauch von Präpositionen und Präpositionalphrasen (PP) (vgl. u. a. Balcı/Kanatlı 2001, Grießhaber 2007, Hufeisen/Gibson 2002, Turgay 2011). Die Mehrheit dieser Untersuchungen widmet sich der Realisierung des regierten Kasus, jedoch bereitet die Wahl der (korrekten) Präposition weitaus größere Probleme (vgl. Grießhaber 2011). Schwierigkeiten der Präpositionswahl sind in unterschiedlichen syntaktischen Funktionen der PP zu beobachten. Bisher fehlen jedoch Erkenntnisse zum Einfluss dieser Funktionen auf den Gebrauch durch Lernende. Die vorliegende Fallstudie[1] präsentiert einen Ansatz, diese For-

1 Die Fallstudie präsentiert vorläufige Ergebnisse aus einem laufenden Dissertationsprojekt.

schungslücke zu schließen. Anknüpfend an Weber (2014, 2015) wird die Verwendung distinkter grammatischer Funktionen von PP durch DaF-Lernende unterschiedlicher Kompetenzniveaus analysiert. Die korpusgestützte Fallstudie untersucht exemplarisch die syntaktischen Funktionen von PP als Objekt und Adverbiale (vgl. Duden 2016: 851f.).

(1) Ich warte auf Ihre Antwort.
‚I'm waiting for your response.'

(2) Ich bin gespannt auf deine Antwort.
‚I'm curious about your response.'

(3) Es gab Kleidung … auf dem Boden.
‚There were clothes … on the floor.'

In der syntaktischen Funktion des Objekts (1)/(2) wird die Präposition vom Verb bzw. Adjektiv spezifiziert und hat ihre primäre Bedeutung verloren (vgl. Duden 2016: 618); sie trägt somit nicht zur Bedeutung der Gesamt-PP bei.[2] In unterschiedlichen Grammatiken wird diese Eigenschaft als konstitutiv für Präpositionalobjekte beschrieben. Die Präposition gilt als „semantisch verblasst" (Eisenberg 2013: 304) oder „semantisch nicht weiter analysierbar" (Helbig/Buscha 2001: 184). Charakteristisch für die Funktion des Präpositionalobjekts ist somit der schwach ausgeprägte semantische Gehalt der Präposition. Im Gegensatz dazu weist die Präposition adverbialer PP (3) eine spezifischere Semantik auf (vgl. Duden 2016: 852) und das unabhängig davon, ob das Adverbiale vom Verb regiert wird oder nicht[3]; Die Präposition trägt hier zur Bedeutung der Gesamt-PP bei.

Die Fallstudie untersucht beispielhaft anhand zweier Präpositionen, ob sich der Unterschied im semantischen Gehalt einer Präposition auf deren Verwendung durch DaF-Lernende auswirkt. Die Forschungsfragen lauten:

— Welche Fehlerhäufigkeiten zeigen sich im Präpositionsgebrauch in den oben genannten syntaktischen Funktionen der PP?
— Welche Fehlertypen im Präpositionsgebrauch zeigen sich in den oben genannten syntaktischen Funktionen der PP?

2 Für einen anderen Ansatz siehe u.a. Zifonun et al. (1997: 1096).
3 Zur Abgrenzung von regierten vs. nicht regierten Adverbialien vgl. z. B. Breindl (2006) oder Zifonun et al. (1997: 2097f.).

2 Fallstudie im Lernerkorpus MERLIN

2.1 Datengrundlage

Die Datengrundlage bildet das deutsche Lernerkorpus MERLIN.[4] Das Korpus enthält authentische Lernertexte, die im Rahmen von standardisierten Sprachtests mit Bezug zum Gemeinsamen Europäischen Referenzrahmen für Sprachen (GeRS) produziert wurden. Die Fallstudie berücksichtigt das GeRS-Gesamtniveau, das auf die Bewertung des produzierten Textes Bezug nimmt (s. dazu Abel et al. 2014: 113f.). Das Korpus verfügt über eine Mehr-Ebenen-Architektur (vgl. Lüdeling et al. 2005), in der u. a. minimale Zielhypothesen (ZH1), d. h. zielsprachliche Rekonstruktionen der Lerneräußerungen, integriert sind (vgl. Lüdeling 2008: 126. Näheres dazu s. MERLIN project 2014: 14ff.), s. Tab.1 zur Illustration. Für den Großteil der Texte liegen Fehlerannotationen vor, u. a. im Bereich Präpositionsgebrauch, der für die Fallstudie zentral ist.

Tabelle 1: Lernertext- und ZH1-Ebene: Beispiel aus dem MERLIN-Korpus.

Lernertext	...	möchte	ich	–	viele	Aktivitäten	teilnehmen	.
ZH1	...	möchte	ich	an	vielen	Aktivitäten	teilnehmen	.

Ausgehend von der Annahme, dass frequente Einheiten in der Zielsprache eines Lerners für den Erwerb der Zielsprache von hoher Bedeutung sind (vgl. Ellis 2002, Tschirner 2006), wurden für die Fallstudie primäre, lokale Präpositionen ausgewählt. Diese sind in der deutschen Sprache hochfrequent (vgl. Eisenberg 2013: 184, Duden 2016: 613) und können sowohl in der semantisch verblassten Verwendungsweise als auch mit eigenständiger Bedeutung gebraucht werden (vgl. Duden 2016: 618). In der lexikalischen Datenbank dlexdb[5] wurden für die lokalen Präpositionen, die die Duden-Grammatik (2016: 616) nennt, die Häufigkeiten ermittelt und die frequenten Präpositionen *an* und *auf* ausgewählt.

Die Abfrage der Präpositionen (inkl. Verschmelzungen) im MERLIN-Korpus erfolgte auf der ZH1-Ebene. Mit Bezug auf die ZH1 lassen sich sprachliche Kontexte ermitteln, in denen eine bestimmte Präposition zielsprachlich gefordert wird; man erhält sowohl Lerneräußerungen, in denen die Präposition korrekt realisiert wurde, als auch solche, in denen zielsprachliche Korrekturen im Bereich Präpositionsgebrauch durchgeführt wurden.

4 Nähere Informationen zum Korpus s. Abel et al. (2014). Das Korpus ist derzeit frei zugänglich unter http://www.merlin-platform.eu.

5 Verfügbar unter http://dlexdb.de/.

2.2 Datenaufbereitung und Datenauswertung

Je Präposition *an* und *auf* (inkl. Verschmelzungen) wurden die ZH1 mit der entsprechenden Lerneräußerung und die Fehlerannotationen im Bereich Präpositionsgebrauch exportiert und für die Analyse aufbereitet. Für jede PP wurden die in der Einleitung beschriebenen syntaktischen Funktionen annotiert: Objekt (mit verblasster Präposition) und Adverbiale (mit nicht verblasster Präposition).[6] Zusätzlich wurden für jede PP in der Annotationskategorie *Fehlertyp* Abweichungen im Präpositionsgebrauch erfasst. Im Korpus fehlende Annotationen im Bereich Präpositionsgebrauch wurden manuell ergänzt und ebenfalls für die Analyse berücksichtigt. In der Fallstudie stehen die Fehlertypen *Tilgung* und *Wahl* im Vordergrund (vgl. Tab. 2).

Tabelle 2: Annotationskategorie *Fehlertyp* im Bereich Präpositionsgebrauch.[7]

Fehlertyp	Beispiel
Tilgung	… möchte ich viele Aktivitäten teilnehmen ‚… I'd like to participate various activities'
Wahl	Dann denke ich um eine kleine Papagei. ‚Then I think at a little parrot.'

Die PP-Instanzen wurden von zwei Annotatorinnen unter Bezug auf ein von der Autorin erstelltes Annotationshandbuch manuell annotiert. Das Inter-Annotator Agreement (IAA) für 100 doppelt annotierte Instanzen betrug $\kappa = 0.9$[8]. Insgesamt wurden 1.053 PP mit *auf* und *an* (inkl. Verschmelzungen) ausgewertet (s. Tab. 3).

Tabelle 3: Übersicht über analysierte PP je GeRS-Gesamtniveau (absolute Zahlen, bereinigt).[9]

	A2/A2+	B1/B1+	B2/B2+	C1	Summe
Objekt	61	147	212	36	456
Adverbiale	157	198	211	31	597
Summe	218	345	423	67	1.053

Je syntaktische Funktion wurden die Anteile der korrekten und inkorrekten Instanzen ermittelt. Instanzen, in denen redundante Präpositionen vorliegen,

6 Dabei wurden u. a. Falsch-Positive aussortiert.

7 Die Beispiele stammen aus dem Korpus. Zu weiteren Fehlertypen in MERLIN siehe Wisniewski et al. (2014: 12).

8 Carletta (1996: 252) spricht von einem κ-Wert von > 0.8 als Repräsentation einer guten Reliabilität.

9 Die GeRS-Gesamtniveaus A1 und C2 wurden aufgrund geringer Instanzenanzahl nicht berücksichtigt.

wurden ebenfalls berücksichtigt, Ellis/Barkhuizen (2005: 79) sprechen hier in Anlehnung an Pica (1984) von der *target-like use analysis*. Die Klassifizierung und Verteilung der Fehlertypen je syntaktischer Funktionen der PP erfolgte durch eine computergestützten Fehleranalyse (vgl. Dagneaux et al. 1998). Die Ermittlung der Fehleranteile sowie der Verteilung der Fehlertypen erfolgte zusätzlich je GeRS-Gesamtniveau (Kontrastive Interlanguage Analyse) (Granger 1996, 2015); diese Analyse ermöglicht es, Aussagen über Entwicklungsverläufe zu treffen.

2.3 Ergebnisse und Diskussion

Die Ergebnisse der Fallstudie zeigen, dass die unterschiedlichen syntaktischen Funktionen der *an-* und der *auf*-PP mit der Realisierung der Präpositionen in diesen Funktionen zusammenhängen (s. Abb. 1).

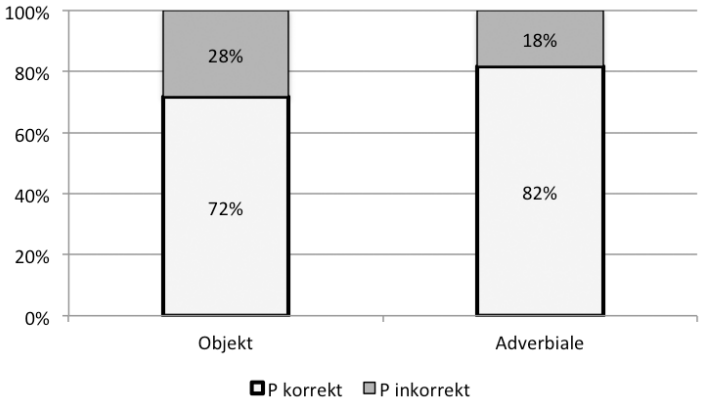


Abbildung 1: Korrekter und inkorrekter Präpositionsgebrauch je syntaktischer Funktion der PP.

Wie in Abb. 1 zu sehen ist, unterscheiden sich die Fehlerhäufigkeiten des Präpositionsgebrauchs bei Objekt und Adverbiale deutlich. Dieser Unterschied ist statistisch signifikant ($\chi^2 = 14{,}304$, df = 1, p = 0,00016).[10] Der Unterschied spiegelt sich ebenfalls in den einzelnen GeRS-Gesamtniveaus der DaF-Lernenden wider (Abb. 2): Die Fehlerhäufigkeiten im Bereich Präpositionsgebrauch unterscheiden sich, in Objekt-PP sind diese konstant höher als bei adverbialen PP. An dieser Stelle muss jedoch auch hervorgehoben werden, dass sich die Fehlerhäufigkeiten beider Funktionen mit Anstieg des GeRS-Niveaus deutlich annähern.

10 Als Signifikanztest wurde der Mehrfelder-χ^2-Test gewählt.

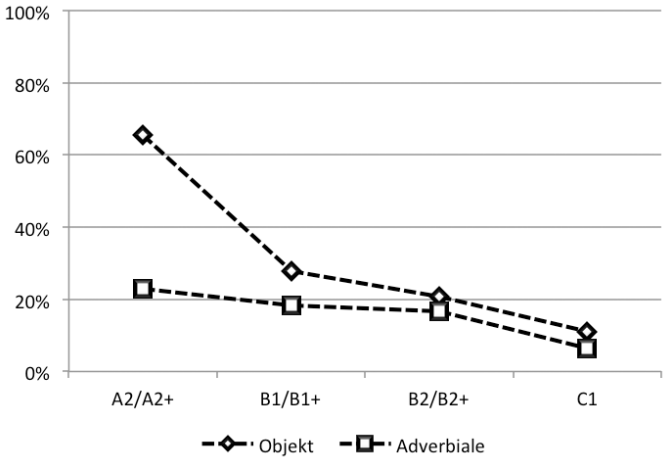


Abbildung 2: Anteile der Fehlerhäufigkeiten je syntaktische Funktion und GeRS-Gesamtniveau.

Die bedeutungsneutralen Präpositionen, die in Objekt-PP enthalten sind, scheinen in der Tat den Präpositionsgebrauch von DaF-Lernenden zu beeinflussen. Die Analyse der fehlerhaften Instanzen zeigt, dass die Fehlertypen *Tilgung* und *Wahl* je syntaktische Funktion der PP unterschiedlich verteilt sind (s. Abb. 3 und 4).[11]

Wie man Abb. 3[12] und 4 entnehmen kann, dominiert der Fehlertyp *Wahl* bei den adverbialen PP fast durchgängig, während sich bei den Objekt-PP mit steigender Sprachkompetenzstufe eine Veränderung der Dominanzreihenfolge vom Fehlertyp *Tilgung* zum Fehlertyp *Wahl* zeigt. Der Anteil des Fehlertyps *Tilgung* bei Objekten mit *an/auf* ist jedoch konstant größer als bei Adverbialien mit *an/auf*. Dieser Unterschied ist statistisch signifikant ($\chi^2 = 17{,}784$, df = 1, p = 2,47e-05). Den Ergebnissen nach könnte der verblasste semantische Gehalt der Präposition in Objekt-PP (gegenüber demjenigen in adverbialen PP) zu größerer Unsicherheit in Bezug auf die Realisierung dieser Präposition führen. Dies zeigt sich vor allem auf den niedrigen und mittleren Kompetenzniveaus, in denen das sprachliche Wissen auf- und ausgebaut wird. Diese Beobachtungen

11 Unter der Kategorie *Rest* sind die Fehlertypen Position und Redundanz zusammengefasst (Näheres dazu s. Wisniewski et al. 2014: 12). Auf diese Fehlertypen wird hier nicht näher eingegangen.

12 An dieser Stelle sei angemerkt, dass im GeRS-Niveau C1 50 % der Fehler darauf zurückgehen, dass eine PP statt einer NP in Objektfunktion realisiert wird (Fehlertyp Redundanz). Insgesamt zeigen sich jedoch auf diesem Niveau (bei beiden PP-Funktionen) sehr wenige Fehler (Objekt: vier, Adverbiale: zwei). Aus diesem Grund sind nur sehr eingeschränkte Aussagen zu diesem GeRS-Niveau möglich.

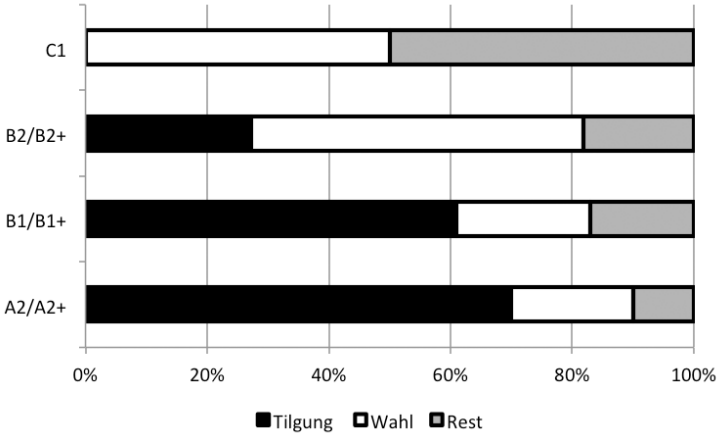


Abbildung 3: Fehlertypen (Präpositionsgebrauch) bei Objekten je GeRS-Gesamtniveau (siehe Anm. 12).

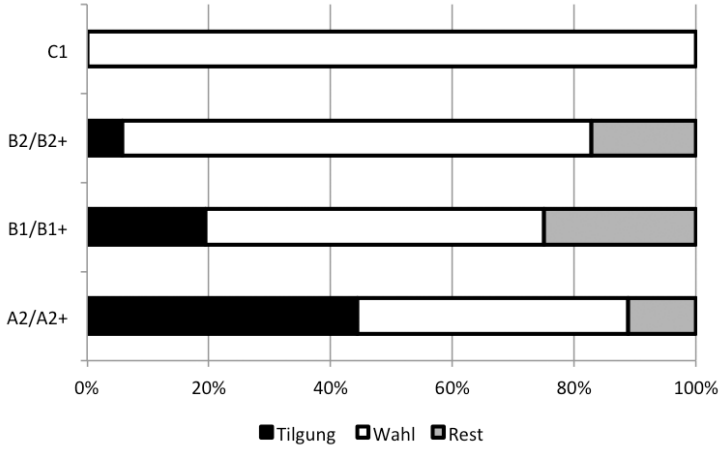


Abbildung 4: Fehlertypen (Präpositionsgebrauch) bei Adverbialien je GeRS-Gesamtniveau.

könnten durch die Annahme erklärt werden, dass sprachliche Einheiten, die nicht salient und notwendig für das Verständnis einer Äußerung sind, erst spät erworben werden (vgl. Ellis 2002: 175). Eine Präposition, die nicht zur Bedeutung der Gesamt-PP beiträgt, wird zunächst vermehrt nicht realisiert. Es kann angenommen werden, dass mit Anstieg der sprachlichen Kompetenz und des sprachlichen Wissens das Wissen um bedeutungsneutrale Präpositionen ebenso auf- und ausgebaut wird. Eine Präposition wird dann seltener getilgt, sondern eher inkorrekt realisiert.

Die Ergebnisse deuten darauf hin, dass die Funktion der PP eine Rolle im DaF-Erwerb von Präpositionen spielt. Es zeigt sich ein Einfluss auf Fehlerhäufigkeit und Fehlertyp. Der semantische Gehalt der Präposition bei Objekten und Adverbialien könnte hierfür verantwortlich sein.

3 Fazit

Die Fallstudie zu den Präpositionen *an* und *auf* im Lernerkorpus MERLIN zeigt, dass der Gebrauch von Präpositionen durch DaF-Lernende von der syntaktischen Funktion der jeweiligen PP beeinflusst wird. Es zeigen sich vor allem Unsicherheiten bei Objekt-PP, in denen die Präposition keinen eindeutigen semantischen Beitrag zur Gesamtbedeutung der PP leistet. Weiterführende Untersuchungen in Lernerkorpora werden durchgeführt, um die Einflussfaktoren auf den Erwerb und Gebrauch bedeutungsneutraler Präpositionen durch DaF-Lernende weiter zu erforschen.

Literatur

Abel, Andrea / Wisniewski, Katrin / Nicolas, Lionel / Boyd, Adriane / Hana, Jirka / Meurers, Detmar (2014): A Trilingual Learner Corpus Illustrating European Reference Levels. In: Ricognizioni – Rivista di Lingue, Letterature e Culture Moderne 2/1, S. 111–126. http://www.ojs.unito.it/index.php/ricognizioni/article/view/702 (27.02.2017).

Balcı, Tahir/Kanatlı, Faik (2001): Das Problem der Kasuswahl nach Wechselpräpositionen. In: Deutsch als Fremdsprache 1, S. 28–30.

Breindl, Eva (2006): Präpositionalphrasen. In: Agel, Vilmos / Eroms, Hans-Werner (Hg.): Dependenz und Valenz/Dependency and Valency. Handbücher zur Sprach- und Kommunikationswissenschaft. 2. Halbband. Berlin/New York: de Gruyter, S. 936–951.

Carletta, Jean (1996): Squibs and Discussions. Assessing Agreement on Classification Tasks: The Kappa Statistic. In: Computational Linguistics 22/2, S. 249–254.

Dagneaux, Estelle/Denness, Sharon / Granger, Sylviane (1998): Computer-aided error analysis. In: System: An International Journal of Educational Technology and Applied Linguistics 26/2, S. 163–174.

dlexDB: Lexikalische Datenbank. Universität Potsdam, Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dlexdb.de/.

Duden (2016): Die Grammatik 9., vollständig überarbeitete und aktualisierte Aufl. Hrsg. von Angelika Wöllstein und der Dudenredaktion. Berlin: Dudenverlag.

Eisenberg, Peter (2013): Grundriss der deutschen Grammatik. Band 2: Der Satz. 4., aktualisierte und überarbeitete Auflage. Stuttgart/Weimar: J. B. Metzler.

Ellis, Nick C. (2002): Frequency Effects in Language Processing. A Review with Implications for Theories of Implicit and Explicit Language Acquisition. In: Studies in Second Language Acquisition 4, S. 143–188.

Ellis, Rod/Barkhuizen, Gary (2005): Analysing Learner Language. Oxford/New York: Oxford University Press.

Granger, Sylviane (1996): From CA to CIA and back: An Integrated Approach to Computerized Bilingual and Learner Corpora. In: Aijmer, Karin/Altenberg, Bengt/Johansson, Mats (Hg.): Languages in Contrast. Text-based Cross-linguistic Studies. Lund: Lund University Press, S. 37–51.

Granger, Sylviane (2015): Contrastive interlanguage analysis: A reappraisal. In: International Journal of Learner Corpus Research 1/1, S. 7–24.

Grießhaber, Wilhelm (2007): „und wir faren in die andere seite" – Der Gebrauch lokaler Präpositionen durch türkische Grundschüler. In: Meng, Katharina & Rehbein, Jochen (Hg.): Kindliche Kommunikation – einsprachig und mehrsprachig. Münster u. a.: Waxmann, S. 371–392.

Grießhaber, Wilhelm (2011): Präpositionen als relationierende Verfahren – Präpositionen vor dem Hintergrund des Türkischen. In: Jahrbuch Deutsch als Fremdsprache 37. München: Iudicium, S. 142-159.

Helbig, Gerhard/Buscha, Joachim (2001): Leitfaden der deutschen Grammatik. Berlin/München u. a.: Langenscheidt.

Hufeisen, Britta/Gibson, Martha (2002): Production of Locative Prepositions by Learners of German as a Second Language. In: Barkowski, Hans/Faistauer, Renate (Hrg): … in Sachen Deutsch als Fremdsprache. Baltmannsweiler: Schneider Verlag Hohengehren, S. 73–90.

Lüdeling, Anke (2008): Mehrdeutigkeiten und Kategorisierung: Probleme bei der Annotation von Lernerkorpora. In: Walter, Maik/Grommes, Patrick (Hg.): Fortgeschrittene Lernervarietäten. Korpuslinguistik und Zweitspracherwerbsforschung. Tübingen: Niemeyer, S. 119–140.

Lüdeling, Anke/Walter, Maik/Kroymann, Emil/Adolphs, Peter (2005): Multilevel error annotation in learner corpora. In: Proceedings from the Corpus Linguistics Conference Series 1/1. http://www.birmingham.ac.uk/research/activity/corpus/publications/conference-archives/2005-conf-e-journal.aspx (27.02.2017).

MERLIN: MERLIN – Multilingual Platform for European Reference Levels: Interlanguage Exploration in Context (Technische Universität, Dresden). http://www.merlin-platform.eu.

MERLIN project (2014): Annotation guidelines. http://www.merlin-platform.eu (27.02.2017).

Pica, Teresa (1984): Methods of Morpheme Quantification: Their Effect on the Interpretation of Second Language Data. In: Studies of Second Language Acquisition 6/1, S. 69–78.

Tschirner, Erwin (2006): Häufigkeitsverteilungen im Deutschen und ihr Einfluss auf den Erwerb des Deutschen als Fremdsprache. In: Corina, Elisa / Marello, Carla / Onesti, Christina (Hg.): Atti del XII Congresso Internationale di Lessicografia. Alessandria: Edizioni dell'Orso, S. 1277–1288.

Turgay, Katharina (2011): Der Erwerb des deutschen Kasus in der Präpositionalphrase. In: Zeitschrift für Germanistische Linguistik 3, S. 24–54.

Weber, Tassja (2014): Verbvalenz und Rektion im Bereich Deutsch als Fremdsprache. Eine korpusgestützte Analyse zweier Verbgruppen (Masterarbeit TU Dortmund). http://merlin-platform.eu/docs/Masterarbeit_Tassja_Weber.pdf (27.02.2017).

Weber, Tassja (2015): Verb Valency and Prepositional Complements in Learner Corpora: A Case Study in the German MERLIN Corpus. In: de Haan, Pieter (Hg.). LCR 2015 Book of Abstracts. Raboud University, S. 164–166. http://www.ru.nl/publish/pages/765127/definitive_book_of_abstracts.pdf (27.02.2017).

Wisniewski, Katrin / Woldt, Claudia / Schöne, Karin / Abel, Andrea / Blaschitz, Verena / Štindlová, Barbara / Vodičková, Kateřina (2014): The MERLIN annotation scheme for the annotation of German, Italian, and Czech learner language. http://www.merlin-platform.eu (27.02.2017).

Zifonun, Gisela/Hoffmann, Ludger / Strecker, Bruno et al. (1997): Grammatik der deutschen Sprache. Band 3. Berlin/New York: de Gruyter.

Christian Lang, Roman Schneider, Karolina Suchowolec

Extracting Specialized Terminology from Linguistic Corpora

Abstract In this paper, we present our approach to automatically extracting German terminology in the domain of grammar using texts from the online information system grammis as our corpus. We analyze existing repositories of German grammatical terminology and develop Part-of-speech patterns for our extraction thereby showing the importance of unigrams in this domain. We contrast the results of the automatic extraction with a manually extracted standard. By comparing the performance of well-known statistical measures, we show how measures based on corpus comparison outperform alternative methods.

Keywords Grammatical terminology, terminological structures, automatic term extraction, grammatical information system

1 Introduction

The information system *grammis* (Schneider and Schwinn 2014) is an online resource on German grammar, hosted by the Institute for the German Language (IDS) in Mannheim. It comprises a wide range of specialist texts on grammatical phenomena of the German language. Additionally, *grammis* offers terminological resources: a dictionary for short reference, and a thesaurus organizing explicit relationships between terminological concepts for the automatic expansion of full-text queries. Established more than a decade ago, the whole system is currently being evaluated and re-designed. As for the current content, we observe that a broad spectrum of grammatical terminology used in the specialist hypertexts is covered neither by the dictionary nor by the thesaurus. We believe and will demonstrate that this coverage can be enhanced by applying automatic term extraction (ATE), i.e. the automatized identification and extraction of terms from domain-specific corpora.

We follow Heylen and De Hertog (2015) by adopting their characterization of a *term* as being part of the "core vocabulary of a specialised domain" (c.f. also Nakagawa and Mori 2002, Kaguera and Umino 1996 among others) which corresponds to German industry standards as defined by DIN2342. However, the classification of a specific entity as *term* (vs. *non-term*) is not a trivial task. Nazar (2016) points out that "in the absence of an intensional definition for the entity *term* researchers must resort to an operational definition" (Nazar 2016: 145), e.g. to a consultation of experts in the domain. In ATE, "the term/non-term categorisation [is] not binary but rather presented as a continuum, in the form of a list of candidates ranked according to a score that represents an estimate of the probability of the candidate being a term" (Nazar 2016: 145). Kageura and Umino (1996: 279f.) point out that the statistical methods used to identify and score term candidates share common assumptions based on the candidates' usage; one of those assumptions appearing more frequently in a specific domain than in general.[1] The quality of an ATE's statistical ranking of candidates can, then, be assessed by the degree to which it coincides with the manual evaluation of the expert.

There has been a substantial amount of research into ATE and its application, however mostly in technological domains (e.g. Nazar 2016, Lossio Ventura et al. 2014, Wermter and Hahn 2005, Frantzi et al. 2000). Zhang et al. (2008) compare different statistical measures applied in automatic term extraction tasks. Their comparative study in the domains of biology and medicine indicates that the domain has an "impact on the performance of ATR[2] algorithms" (Zhang et al. 2008: 2111). They also note that "[…] evaluation in other kinds of domains, notably less technical ones, have been lacking" (Zhang et al. 2008: 2109).

In this paper, we present our approach to extract relevant terminology in the domain of German grammar. As there is – to our knowledge – no evaluation study for this domain, we focus on a comparison of different algorithms. Hence, we implement an array of well-established statistical measures used in automatic term extraction tasks with an emphasis on contrasting corpus comparing measures with alternative measures. We evaluate the performance of the extraction algorithms by comparing the ATE's results to a standard manually extracted by a terminology/linguistics expert (MTE).

1 Kageura and Umino (1996: 280) also point out that while those assumptions seem reasonable, „the task of proper theorization is yet to be carried out."

2 Zhang et al. (2008) use the term Automatic Term Recognition (ATR) instead of Automatic Term Extraction (ATE).

2 Corpus

Our test set of *grammis* texts constitutes a corpus of 2,491 documents with a total of 1.2 million tokens and 44,000 types. Contents range from concise descriptions to more detailed discussions. From a technical point of view, all primary data and meta-data is coded within semi-structured XML instances that are composed of semantic markup elements ("title", "subtitle", "literature" etc.). As common in linguistic texts, most of the documents contain natural language example sentences for illustration purposes. These sentences, mostly taken from newspaper articles, are not consistently identified by semantic markup. This results in a substantial number of non-domain specific words which ATE has to handle.

3 Method

We start with standard linguistic preprocessing – applying TreeTagger (Schmid 1995), we assign Part-of-speech tags (POS) and stem the words in the corpora. After that, we apply three filters in order to block undesired candidates from extraction: the first filter exploits the semantic markup of the XML instances. In particular, it excludes bibliographical references and example sentences if they are marked as such. The second – statistical – filter is based on a comparison of our target corpus with a general domain reference corpus (see 3.2). A term candidate is eligible for extraction only if its relative frequency is higher in the specialized target corpus than in a general domain reference corpus (see Gelbukh et al. 2010). The statistical filter is implemented to minimize the amount of noise that is introduced by the non-terminological example sentences. No absolute frequency threshold is applied.[3] The third filter is based on POS patterns as described in 3.1. All candidates that satisfy the POS filter, the relative frequency threshold, and the semantic markup-filter are extracted from our target corpus.[4] They are subsequently ranked by the algorithms described in 3.2.

3 The manually extracted standard (see 4) includes a total of 67 hapax legomena with a frequency of 1, e.g. *Pseudocleft-Satz* ('pseudo cleft sentence').

4 Coordinated composites are a special challenge for extraction. Coordinated nouns share a morpheme that is omitted in one of them, e.g.: *Ereignis- und Betrachtzeit* ('event time and focus time'). Both, *Ereigniszeit* and *Betrachtzeit* are key terms, whereas the coordination is not. We extract the coordination and treat both coordinated elements as unigrams.

3.1 Linguistic Filter – POS Patterns

Justeson and Katz (1995) propose POS patterns for terminology extraction in English by analyzing dictionaries of different technical domains. The benefit of applying POS filters is the improvement of precision. The drawback is a potentially reduced recall. In order to minimize the risk of a reduced recall based on too narrow POS filters, we analyze the prevalent POS patterns of German grammatical terms in the above-mentioned *grammis* thesaurus and in the online version of the alphabetic index of *Duden – die Grammatik* (Duden 2017). The analysis of a total of 2,984 terms shows that 82 % of them are either nominal or adjectival unigrams, while only 15 % are bigrams of an adjective and a noun. These results contrast with Justeson and Katz (1995) who find that "the majority of technical terms do consist of more than one word" (Justeson and Katz 1995: 9); this observation, however, is based on English dictionaries in technical domains.

Our POS filter incorporates the following patterns that represent 99 % of the terms analyzed: N, A, AN, NN, N Prep N, N Det N, (V), A A N.[5]

3.2 Ranking Candidates

In order to rank the extracted candidates, we compare a series of well-established statistical measures that have been used in similar automatic term extraction tasks (see Heylen and De Hertog 2015 or Zhang et al. 2008 for an overview).[6] The implemented measures fall into one of two categories: measures based on corpus comparison and measures not based on corpus comparison. For the first type, our target corpus is compared to a randomly extracted sample from DeReKo (German Reference Corpus; Kupietz and Keibel 2009). It covers various text types and genres, and contains approx. 970,000 tokens and 80,000 types. In this group

5 **N**: nouns, proper names, numbers; **A**: adjectives, attributive and predicative; **Prep**: prepositions, **Det**: determiners, **V**: verbs. However, we exclude verbs from the extraction. With a share of a mere 0.34 % of the analyzed grammatical terms and a share of 11 % of the words in our target corpus, the inclusion of verbs would have increased noise for a minor improvement of recall.

6 Some of the measures we implemented are also used in the extraction of keywords (c.f. Heylen and De Hertog 2015: 219, also Kageura and Umino (1996) for a discussion of the close relation between the two fields). The measures we implemented have been used to extract terms in other (technological) domains, for example: **LL** by Gelbukh et al. (2010) for computer science, **Weird** by Gillam et al. (2007) for nanotechnology, **C-value** by Frantzi et al (2000) for medicine, **P-Mod** Wermter and Hahn (2005) for biomedicine. **TFIDF**, while prototypically applied in keyword extraction, is used by Zhang et al. (2008) as a baseline in their comparative study.

we implement the following measures: *Log-Likelihood based distance* – **LL** (Dunning 1993), *Simple Math* (with an *add-N parameter* of 10) – **SM_10** (Kilgarriff 2009) and *Weirdness* – **Weird** (Ahmad et al. 1999). All these measures evaluate a candidate's termhood (in the sense of Kageura and Umino 1996) and are based on the presumption that "terms are by definition domain-specific, and as a consequence are hypothesised to occur more frequently in their proper domain than they do in other domains or in general language use" (Heylen and De Hertog 2015: 219). While comparing the corpora, bigrams and trigrams are treated the same way as unigrams. Since bigrams and trigrams are generally less frequent, they are ranked lower in comparison to unigrams. For terms spanning more than one word, this is a crucial point in the analysis. The C-value and P-Mod measures (see below) are one way of incorporating information about the frequency of multi-word units and their relationship to the frequencies of shorter multi-word units contained in them.

The second type of measures is not based on corpus comparison. We implement three measures of this type: first, **TFIDF** (Spärck Jones 1972), which is widely used in text mining. TFIDF weighs a candidate's frequency in the corpus with its document frequency. Second, Frantzi et al.'s **C-value** (2000), which is based on frequency, and takes into consideration a candidate's likelihood of being nested in a construction. We use a modified version to account for unigrams (Lossio-Ventura et al. 2014). In the third place, we implement Wertmer and Hahn's paradigmatic modifiability, **P-Mod** (Wermter and Hahn 2005), also in a modified version to account for unigrams. Both C-value and P-Mod are hybrid approaches that combine a candidate's unithood and termhood (in the sense of Kageura and Umino 1996) and were both originally designed to identify multigram terms. In a final step, we also implement **t-value** to assess the unithood of multigrams and a distance metric based on longest common subsequence to detect spelling variants among the candidates. For calculating bonuses, we use semantic markups from the original XML files. Candidates receive a bonus of 30 % or 10 % if they are mentioned in a title or a subtitle respectively.

4 Results and Discussion

To evaluate our ATE results, we ask a linguistic terminologist to perform a manual terminology extraction from a randomly chosen subset of 120 out of the 2,491 documents in the corpus. The expert is asked to extract all linguistic terms regardless of structure, i.e., without POS-filtering. The results of this manual extraction serve as a gold standard for the quality of our ATE. We choose this design over a manual evaluation of the term candidates identified by the ATE as we want to prevent a bias towards parameters inherent to the ATE.

The manual extraction results in a list of 1,001 terms.[7] A large majority of 98 % are nouns, adjectives and their combination. 82.6 % of the manually extracted terms are unigrams, which corresponds to our analysis of existing repositories of German grammar described in 3.1. 948 of these standard terms are also found by ATE. With a total of 5,314 ATE candidates, this means a recall of 94.7 % with an overall precision[8] of 17.8 %.

The imperfect recall score cannot be attributed to the narrow POS-filter. Six terms in the standard are not in the scope of the ATE's POS filter; five of them are verbs. We observe that nominal and/or adjectival equivalents of all those verbs are retrieved by ATE. The main reason (27 %) for the imperfect recall score is a higher relative frequency in the general domain reference.

Regarding precision, the analysis of the top-ranked candidates missing in the standard shows at least five obvious key terms such as *flexion*, *outside field*, *phonological*, *unmarked* and *unstressed*. Besides, a candidate's spelling variants are sometimes treated differently by the expert: e.g., *Aufforderungsmodus* ('prompt mode') was deemed a term, whereas *Aufforderungs-Modus* was not. We attribute this to performance errors by the expert rather than to lack of expertise. In any case, this is a strong argument for always combining manual and automatic term extraction: the major advantage of manual extraction is the specialized knowledge of the expert; the brute force of ATE and its being based on objective corpus evidence can compensate for possible performance errors by the expert.

We further evaluate the precision of the ATE by ranking the candidates according to the implemented measures described in 3.2. Table 1 shows the ATE's precision for all implemented measures at various cutoffs, thus for the top i ranked candidates each. The results indicate that the precision of corpus-comparing measures is generally higher than the measures based on the target corpus only; *Weirdness* demonstrates the highest precision.

7 We retrospectively excluded a total of 28 terms from the standard. This was done either because of typos or because their exact form was not found in the documents. This applies primarily to complex NPs such as *local and temporal adverbials*. The expert extracted both *local adverbials* and *temporal adverbials*, even though the exact string *local adverbials* is not present in the text.

8 Recall is the fraction of terms that were successfully extracted: $\mathrm{R} = \frac{\text{correctly extracted terms}}{\text{all standard terms}}$.

Precision is the fraction of extracted candidates that are terms: $\mathrm{P} = \frac{\text{correctly extracted terms}}{\text{all extracted candidates}}$.

Table 1: Precision of ATE.

	Top i Ranked Candidates Evaluated			
Ranking Method	i = 50	i = 100	i = 500	i = 1000
Freq	56.0 %	60.0 %	45.2 %	38.5 %
TFIDF	76.0 %	68.0 %	51.8 %	42.9 %
Weird	96.0 %	88.0 %	67.6 %	51.4 %
SM_10	90.0 %	77.0 %	57.0 %	44.6 %
LL	78.0 %	75.0 %	57.6 %	45.4 %
C-value	66.0 %	68.0 %	51.0 %	39.9 %
P-Mod	58.0 %	62.0 %	46.6 %	38.3 %

Taking recall into account, Figure 1 displays precision-recall curves for all implemented measures. Increasing recall, the decrease in precision is slower for *Weirdness'* than for the other measures.

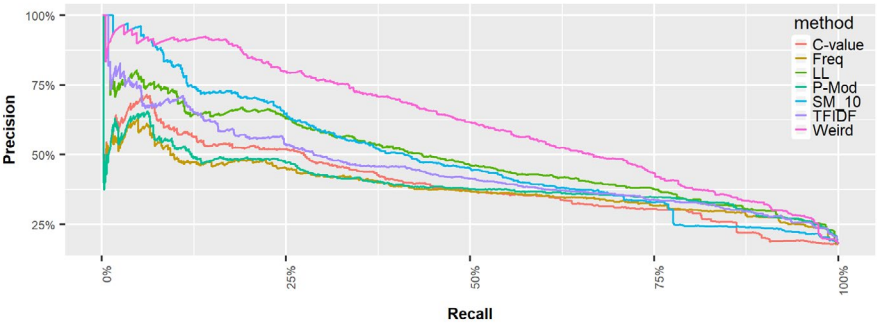


Figure 1: Precision/Recall graph.

As another metric to evaluate the ranking measures, we calculate the *Average Precision* (*AvP*) (Su et al. 2015) which is defined as:

$$\sum_{i=1}^{N} P(i)\Delta R(i)$$

In this formula, N represents the total number of candidates, P(i) is the precision at a cutoff of i candidates and ΔR(i) is the change in recall between cutoff i-1 and i. The AvP score is higher the more actual terms are among the higher ranked candidates. Figure 2 shows the AvP values for the examined measures:

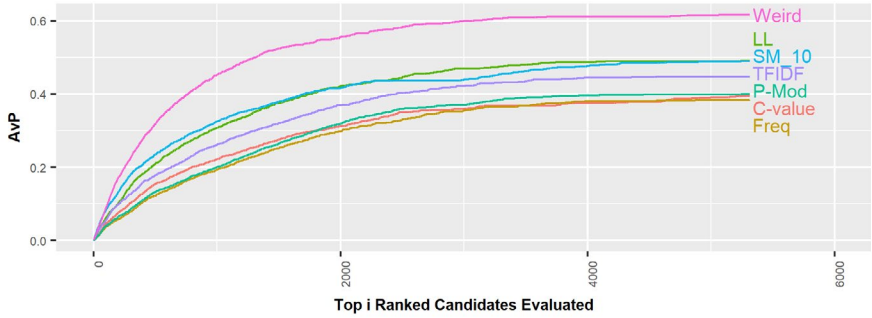


Figure 2: Average Precision (AvP).

Overall, the *Weirdness* measure shows the best performance; compared to the other measures, higher ranked candidates are more likely to be terms.[9] Measures that are based on corpus comparison outperform those that are based on the target corpus only. We attribute this result to the subset of high frequency candidates which are part of the general domain, e.g. *difference*, *example*. The comparison with a general language corpus results in a lower ranking of those candidates.[10] Finally, due to the high proportion of unigrams among the terms manually extracted by the expert, the algorithms that were designed to identify multigram terms show a weaker performance.

5 Concluding Remarks

We presented our approach to extract German grammatical terminology from linguistic corpora, and compared the performance of different ATE methods in this domain. The results indicate that corpus-comparing methods perform better than measures that are not based on corpus comparison. We showed the importance of unigrams in the domain of German grammar by analyzing both existing terminology repositories and the results of the manual extraction by an expert. This result contrasts with the prevalence of multigram terms in technical domains as stated by Nakagawa and Mori (2002) or Justeson and Katz (1999). The tendency towards shorter terms can be interpreted as characteristic for the domain of grammar, confirming Frantzi et al.'s (2000) observation that terms

9 In Zhang et al. (2008) *Weirdness* outperformed *TFIDF* and *C-value* when applied to the Wikipedia Corpus, however performed worse when applied to the life science corpus Genia.

10 Six of the ten most frequent candidates are words of the general domain. *C-value* and *P-Mod* rank five of them in their top ten.

tend to be shorter in arts compared to science and technology. Furthermore, German word formation allows for complex compound-unigrams that correspond to multiword units in English.

References

Ahmad, Khurshid, Lee Gillam and Lena Tostevin. 1999. University of Surrey Participation in TREC8: Weirdness Indexing for Logical Document Extrapolation, Retrieval (WILDER). In Ellen Voorhees and Donna Harman (eds.), *NIST Special Publication 500-246: The Eighth Text Retrieval Conference (TREC-8)*, 717—724. Gaithersburg, MA.

DIN 2342. 2011. *Begriffe der Terminologielehre*. Berlin: Beuth Verlag.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Journal of Computational Linguistics — Special Issue on Using Large Corpora*, 19(1): 61—74.

Duden. 2017. Grammatische Fachausdrücke. http://www.duden.de/sprachwissen/sprachratgeber/Grammatische-Fachausdrucke (February 02, 2018).

Frantzi, Katerina, Sophia Ananiadou and Hideki Mima. 2000. Automatic Recognition of Multi-Word Terms: the C-value/NC-value Method. *International Journal on Digital Libraries* 3(2): 115—130.

Gelbukh, Alexander, Grigori Sidorov, Eduardo Lavin-Villa and Liliana Chanona-Hernandez. 2010. Automatic Term Extraction Using Log-Likelihood Based Comparison with General Reference Corpus. In *Proceedings of the 15th International Conference on Application of Natural Language Processing to Information Systems, NLDB 2010*, 248—255. Berlin/Heidelberg: Springer.

Gillam, Lee, Mariam Tariq and Kurshid Ahmad. 2007. Terminology and the Construction of Ontology. In Fidelia Ibekwe-SanJuan, Anne Condamines and M. Teresa Cabré Castellví (eds.), *Application-Driven Terminology Engineering*. Benjamins Current Topics 2: 49—73.

Heylen, Kris and Dirk De Hertog. 2015. Automatic Term Extraction. In Hendrik J. Kockaert and Frieda Steurs (eds.), *Handbook of Terminology. Volume 1*, 203—221. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Justeson, John S. and Slava M. Katz. 1995. Technical Terminology: some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering* 1(1): 9—27.

Kageura, Kyo and Bin Umino. 1996. Methods of Automatic Term Recognition: A Review. *Terminology* 3(2): 259—289.

Kilgarriff, Adam. 2009. Simple Maths for Keywords. In Michaela Mahlberg, Victorina González-Díaz and Catherine Smith (eds.), *Proceedings of Corpus Linguistics Conference CL2009*, University of Liverpool, UK, July 2009.

Kupietz, Marc and Holger Keibel. 2009. The Mannheim German Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research. In Makoto Minegishi and Yuji Kawaguchi (eds.), *Working Papers in Corpus-based Linguistics and Language Education*, No. 3., 53—59. Tokyo: Tokyo University of Foreign Studies (TUFS).

Lossio Ventura, Juan Antonio, Clement Jonquet, Mathieu Roche and Maguelonne Teisseire. 2014. Biomedical Terminology Extraction: A new Combination of Statistical and Web Mining Approaches. *Proceedings of Journées Internationales d'Analyse Statistique des Données Textuelles (JADT 2014)*, Paris, France.

Nakagawa, Hiroshi and Tatsunori Mori. 2002. A Simple but Powerful Automatic Term Extraction Method. *Proceedings of the Second International Workshop on Computational Terminology*, Stroudsburg, PA, USA: ACL: 1—7.

Nazar, Rogelio. 2016. Distributional Analysis Applied to Terminology Extraction. First Results from the Domain of Psychiatry in Spanish. *Terminology* 22(2): 141—170.

Schmid, Helmut. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*, Dublin, Ireland: 1—9.

Schneider, Roman and Horst Schwinn. 2014. Hypertext, Wissensnetz und Datenbank: Die Webinformationssysteme grammis und ProGr@mm. In Franz Josef Berens and Melanie Steinle (eds.). *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*, 337—346. Mannheim: IDS.

Spärck Jones, Karen. 1972. A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation* 28(1): 11—21.

Su Wanhua, Yan Yuan and Mu Zhu. 2015. <http://dx.doi.org/10.1145/2808194.2809481>. *Proceedings of the ACM SIGIR 2015 International Conference on the Theory of Information Retrieval*: 349—352.

Wermter, Joachim and Udo Hahn. 2005. Paradigmatic Modifiability Statistics for the Extraction of Complex Multi-Word Terms. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*: 843—850.

Zhang, Ziqi, José Iria, Christopher Brewster and Fabio Ciravegna. 2008. A Comparative Evaluation of Term Recognition Algorithms. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*: 2108—2113. Marrakech: ELRA.

Beatrix Busse, Kirsten Gather, Ingo Kleiber

Assessing the Connections between English Grammarians of the Nineteenth Century – A Corpus-Based Network Analysis

Abstract Linguistic studies of nineteenth-century British grammar books are still scarce despite essential changes in the genre during the nineteenth century, such as the decline of so-called prescriptive grammar writing. Since grammarians often use references to other authors to criticise the seemingly inadequate works of predecessors and contemporaries, our study investigates the scholarly network of grammarians' references in a corpus of nineteenth-century English grammars. We particularly focus on the transition from prescriptive to descriptive grammar writing, showing that this paradigmatic turn in the genre is reflected both in the network of grammarians' references and in the usage of terms like prescriptive and descriptive in the grammars.
Our study is part of the *HeidelGram* project, which combines methods from corpus-based diachronic linguistics and network analysis with the aims of offering new perspectives on (meta-)linguistic developments and to reassess well-established assumptions on the history of the genre *grammar*.

Keywords English grammar, corpus, nineteenth century, prescriptive, network analysis, references

1 Introduction

Systematic and comprehensive linguistic studies of nineteenth-century British grammar books are scarce[1] although the nineteenth century is often seen as a turning point in English grammar writing, in particular due to the assumed

1 Anderwald, who studied different aspects of verb morphology and syntax in nineteenth-century English and American grammars (e.g. Anderwald 2014, 2016), also considers this area of investigation "still a gap" (Anderwald 2016: 3).

paradigm shift from prescriptive to predominantly descriptive grammars (e.g. Finegan 1998: 559ff). In particular, authors' references to other grammarians show that new as well as outdated approaches to grammar writing were discussed extensively, often with the aim to justify one's own and better contribution.

„Onomastic" references, that is, references to authors' names, form an important indicator of how nineteenth-century grammarians interacted with each other. Therefore, it makes sense to examine the connection between these references and different linguistic approaches to grammar writing. In the present pilot study, which is part of the *HeidelGram* project[2], we focussed on the turn away from the prescriptive tradition towards a new, descriptive approach to grammar. We built and analysed a network of grammarians' references on the basis of a corpus of nineteenth-century British grammar books, thus combining methods from historical corpus linguistics and network analysis. Additionally, the frequency analysis of the terms *prescriptive/prescription* and *descriptive* is used to illustrate whether the lexico-grammatical inventories of the nineteenth-century grammars under investigation also point to the assumed changes in the genre.

2 Pilot Study

The pilot corpus of nineteenth-century grammar books compiled for this study contains 40 texts, which amount to ca. 2.6 million words. The choice of grammar books was guided by several criteria, such as the popularity and distribution of the grammars (see, for instance, Michael 1987, Görlach 1998), and their variety in function, audience, and text type.

2.1 Scholarly Network Studies

Most commonly, network-analytic approaches are used to examine the relations between people, groups, or organisations. In contrast to such social networks, scholarly networks can feature both social ties as well as cultural ties „beyond the boundaries of personal acquaintanceship" (White 2011: 271). This kind of non-social relationship can often be observed when scholars cite other scholars that are personally unknown to them.

In this study, the relationships between grammar books are assessed in the form of a network of grammar books and the authors that are referenced in them. The two kinds of nodes in this scholarly network are the nineteenth-century grammar books in which references to other grammarians are found, and

2 See http://heidelgram.uni-heidelberg.de for details.

authors' last names, which are used as search terms. The search terms were compiled by collecting the last names of those who are considered to be the most popular and influential grammarians of the sixteenth to nineteenth centuries (see, for instance, grammarians mentioned in Finegan 1998, Michael 1987).

2.2 Automated Network Generation

There are hardly any reliable, machine-readable versions of nineteenth-century grammar books available, and manually generating complex author-text networks of this size is not feasible. Therefore, an automated network-generation process based on threshold-based gestalt pattern matching[3] and manual elimination of false positives was developed.

First, the *pdf*-scans of grammar books were digitised using the *Google*-maintained *Tesseract* OCR software. Despite rather acceptable text recognition results, the OCR software is susceptible to producing output containing misreadings. Hence, the resulting text files were cleaned up using *HGAutoFix* by applying a pre-defined set of corrective rules, e.g. normalising punctuation and spelling, significantly enhancing the quality of the data.

The data then was passed into *HGSimpleCorpusNetwork*[4], which created a document-term matrix, a list of concordances, and the respective network graph in GraphML format from the given set of text files and the list of search terms. To account for OCR-corrupted data, the search algorithm supports approximate string matching utilising Levenshtein distances and gestalt pattern matching with user-defined thresholds (0.8). Due to this error-tolerant, but approximate approach, the resulting data needed to be manually reviewed and false positives had to be removed.

Gephi was then used for exploratory data analyses. The network was visualised as a circular graph utilising the *layout_in_circle* graph layout of the *igraph*[5] network-analysis package in R. The size of the *grammar* nodes was derived from the number of tokens divided by 10,000. The thickness of the edges was kept proportionate to the number of references.

3 Our software *HGSimpleCorpusNetwork* utilises the Python *difflib* implementation of the Ratcliff/Obershelp pattern-recognition algorithm (cf. https://docs.python.org/3/library/difflib.html).

4 The software is freely available on GitHub: https://github.com/heidelgram/HGSimpleCorpusNetwork.

5 See http://igraph.org.

2.3 Results

This section sums up main results of network and frequency analyses, focussing on the temporal distribution of the references and testing established knowledge about the transition from prescriptive to descriptive grammar writing in the nineteenth-century.

The search for grammarians' last names in the 40 grammar books led to a list of 1,518 references to other grammarians. Although search terms comprised the allegedly most popular and influential grammarians of their time, many earlier grammarians did not play a role in nineteenth-century grammar writing any more, apart from Ben Jonson's grammar (1640), which was still considered a valuable source with regard to Early Modern English pronunciation. Robert Lowth and Lindley Murray, who are usually considered the major and most popular presciptivists (e.g. Beal 2004: 89f, Auer 2008: 58), are among the most frequently referenced grammarians in the corpus (see Busse, Gather, Kleiber: forthcoming). References to them, however, did not necessarily imply agreement, but are rather a means of expressing criticism.

Figure 1 illustrates the references to grammarians from 1800 to 1900 as a network. This visualisation, resembling small-world networks, was chosen because it is particularly well-suited to show the temporal distribution of references, and the most often referenced as well as referencing authors in one graph. The circles in the upper half of the network are the referenced search terms, i.e. those last names of grammarians that were referred to in at least one nineteenth-century grammar book. The squares below represent the nineteenth-century grammars. References to the search terms are visualised by edges of different sizes, the size corresponding to the number of references made.

Figure 1 shows that most of the citations refer to eighteenth- and nineteenth-century grammarians, in particular to Lowth, L. Murray, and Tooke. Most of the references stem from the grammars by Crombie (1802), Cramp (1838) and Gerald Murray (1847). The network graph indicates a break in dealing with other grammarians around 1850. While before authors often referred to 18th- and early nineteenth-century grammarians, similar references become very rare in the second half of the nineteenth century and authors often focus on their contemporaries. Considering which authors were referenced by grammarians of the first half of the nineteenth century, there is a turn away from the occupation with prescriptive grammar authors like Lowth and Lindley Murray.

With regard to the change in focus of grammar writing around 1850 that can be assumed from Table 1, the question arose whether this transition from prescriptive to descriptive grammar writing co-occurs with lexemes which refer to the respective concepts, i.e. *prescriptive*/*prescription* and *descriptive*, in the nineteenth-century grammars. According to historical linguists, prescriptive

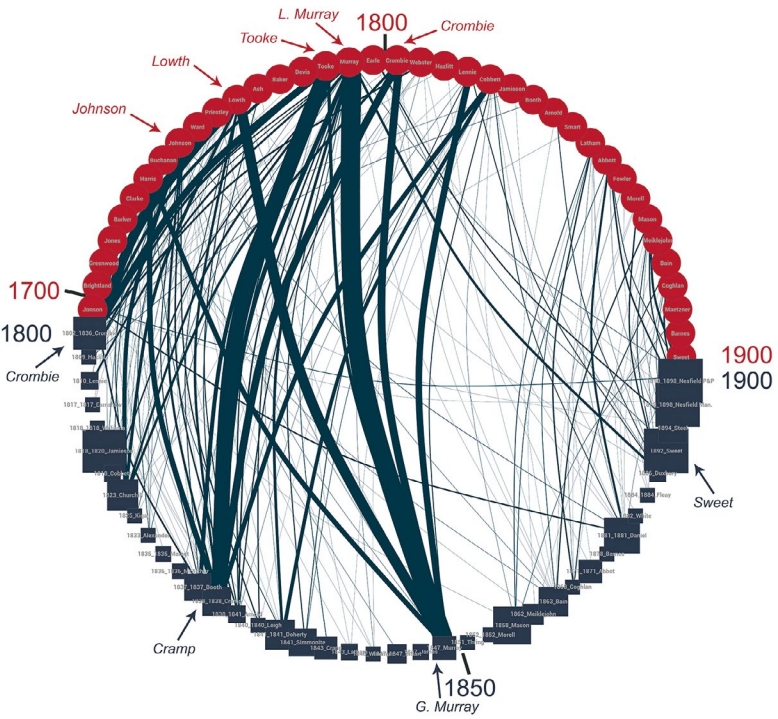


Figure 1: Grammarians' references to other grammarians in the corpus (red circles = referenced authors; blue squares = nineteenth-century grammars).

grammar writing emerges in the second half of the 18th century (e.g. Tieken-Boon van Ostade 2008: 6) and was at its height in the first half of the nineteenth century (Dekeyser 1975: 266). Frequency analyses of the grammars' lexical inventories show that indeed the terms *prescriptive* or *prescription* are used sporadically by three authors in the first half of the nineteenth century, but not after 1850. This corresponds to the findings in Figure 1, which contains hardly any references back to prescriptivists like Lowth and Murray after 1850. Henry Sweet coins the term *descriptive grammar* in his grammar (1892/98), which is usually considered the first important descriptive and historical grammar (e.g. Beal 2004: 115).

It should, however, be noted that these findings only indicate the first occurrences of certain terms, but not of their concepts, and that prescriptive aspects are likely to manifest not necessarily in the terms *prescriptive* or *prescription*, but in expressions such as *bad English*, *improper grammar*, and *solecism*. The diachronic analysis of changing terms and related concepts in historical grammar books (i.e. form to function and function to form) is a future task of the *HeidelGram* project.

Table 1: Corpus of Nineteenth-century Grammar Books

Author	**Year**	**Title**
Abbott, Edwin A.	1871	*English Lessons for English People*
Alexander, Levy	1833	*The Young Lady and Gentleman's Guide to the Grammar of the English Language in Verse*
Arnold, Thomas K.	1838	*An English Grammar for Classical Schools*
Bain, Alexander	1863	*An English Grammar*
Barnes, William	1878	*An Outline of English Speech-Craft*
Booth, David	1837	*The Principles of English Grammar*
Churchill, T.O.	1823	*A New Grammar of the English Language*
Cobbett, William	1818	*Grammar of the English Language, in a Series of Letters*
Coghlan, John	1868	*Reformed English Grammar*
Cramp, William	1838	*The Philosophy of Language*
Crane, George	1843	*The Principles of Language; Exemplified in a Practical English Grammar*
Crombie, Alexander	1802	*The Etymology and Syntax of the English Language, Explained and Illustrated*
Daniel, Rev. Evan	1881	*The Grammar, History and Derivation of the English Language*
Doherty, Hugh	1841	*An Introduction to English Grammar, on Universal Principles*
Duxbury, C.	1886	*A New English Grammar of School Grammars*
Earnshaw, Christopher	1817	*The Grammatical Remembrancer*
Fleay, Frederick G.	1884	*The logical English grammar*
Hazlitt, William	1809	*A New and Improved Grammar of the English Tongue*
James, J.H.	1847	*The Elements of Grammar, according to Dr. Becker's System*
Jamieson, Alexander	1818	*A grammar of rhetoric and polite literature*
Kigan, John	1825	*A Practical English Grammar, agreeably to a new System*
Latham, Robert G.	1843	*An Elementary English Grammar*
Leigh, Percival	1840	*The Comic English Grammar*
Lennie, William	1810	*The principles of English grammar briefly defined, and neatly arranged*
Marcet, Jane	1835	*Mary's Grammar*
Mason, C. P.	1858	*English Grammar; including the Principles of Grammatical Analysis*
McArthur, Alexander	1836	*An outline of English grammar for the use of schools*
Meiklejohn, John	1862–66	*An Easy English Grammar for Beginners*
Morell, John D.	1852	*The analysis of sentences explained and systematised*
Murray, Gerald	1847	*The Reformed Grammar, or Philosophical Test of English Composition*
Nesfield, John C.	1898a	*English Grammar Past and Present*

Table 1: Corpus of Nineteenth-century Grammar Books (continued).

Author	Year	Title
Nesfield, John C.	1898b	*Manual on English Grammar and Composition*
Simmonite, Wiliam J.	1841	*The Practical Self-teaching Grammar of the English Language*
Smart, Benjamin H.	1847	*Grammar on its True Basis*
Steel, G.	1894	*An English grammar and analysis for students and young teachers*
Sweet, Henry	1892/98	*A New English Grammar: logical and historical*
Thring, Rev. Edward	1851	*The Elements of Grammar Taught in English*
White, Frederick Averne	1882	*English Grammar*
Williams, David	1818	*The catechism of English grammar*
Wiseman, Thomas J.	1846	*A School Grammar of the English Language*

Table 2: Other grammar books

Author	Year	Title
Jonson, Ben	1640	*The English Grammar*
Lowth, Robert	1762	*A Short Introduction to English Grammar with Critical Notes*
Murray, Lindley	1795	*English Grammar Adapted to the Different Classes of Learners*

3 Summary and Conclusion

In British grammar writing, the nineteenth century is usually considered as a transition period from the prescriptive tradition to a new, descriptive approach to grammar.

The present pilot study investigated the scholarly network of nineteenth-century grammarians, as manifested by their references to other grammarians, focussing on the move away from the occupation with so-called prescriptive grammar writing. The network revealed a substantial change around 1850, indicating that grammars after 1850 seem to become more and more independent from the prescriptive tradition, and from the prescriptivists Lowth and L. Murray in particular. Frequency analyses showed that the terms *prescriptive* or *prescription* are indeed used sporadically in the first half of the nineteenth century, usually combined with a critical remark on the rigidity of prescriptive grammar writing, and that *descriptive* in connection with grammar writing was coined by Henry Sweet in the 1890s.

For two reasons, however, results should be treated with caution. As mentioned in 2.3., the findings only give evidence about first occurrences of linguistic terms, not about their underlying concepts. Follow-up studies within the

HeidelGram project will examine the development of linguistic terminology and concepts in 16th-to nineteenth-century grammars.

The other reason relates to the quality of the corpus data. To account for OCR-corrupted data, it makes sense to work with a low pattern-matching threshold, despite the higher effort of manual correction, in order not to miss results. The present pilot study shows that although the data have not yet been revised manually, significant results could nevertheless be obtained, but caution is advised.

References

Anderwald, Lieselotte. 2014. The decline of the BE-perfect, linguistic relativity, and grammar writing in the 19th century. In Marianne Hundt (ed.), *Late Modern English Syntax*, 13–37. Cambridge: Cambridge University Press.

Anderwald, Lieselotte. 2016. *Language Between Description and Prescription: Verbs and Verb Categories in Nineteenth-Century Grammars of English*. Oxford: Oxford University Press.

Auer, Anita. 2008. Eighteenth-century grammars and book catalogues. In Ingrid Tieken-Boon van Ostade (ed.), *Grammars, Grammarians and Grammar Writing in Eighteenth-Century England*, 57–75. Berlin and New York: Mouton de Gruyter.

Beal, Joan C. 2004. *English in Modern Times. 1700–1945*. London: Arnold.

Busse, Beatrix, Kirsten Gather and Ingo Kleiber. forthcoming. Paradigm Shifts in 19th-Century British Grammar Writing – a Network of Texts and Authors. In Claudia Claridge and Birte Bös (eds.), *Norms and Conventions*. Amsterdam/Philadelphia: John Benjamins.

Dekeyser, Xavier. 1975. *Number and Case Relations in 19th Century British English: A Comparative Study of Grammar and Usage*. Bibliotheca Linguistica. Antwerpen et al.: De Nederlandsche Boekhandel.

Finegan, Edward. 1998. English Grammar and Usage. In Suzanne Romaine (ed.), *The Cambridge History of the English Language Vol. IV: 1776–1997*, 536–588. Cambridge: Cambridge University Press.

Görlach, Manfred. 1998. *An Annotated Bibliography of Nineteenth-Century Grammars of English*. Amsterdam/Philadelphia: John Benjamins.

Michael, Ian. 1987. *The Teaching of English*. Cambridge: Cambridge University Press.

Tieken-Boon van Ostade, Ingrid (ed.), 2008. *Grammars, Grammarians and Grammar-Writing in Eighteenth-century England*. Berlin/New York: Mouton de Gruyter.

White, Howard D. 2011. Scientific and Scholarly Networks. In John Scott and Peter J. Carrington (eds.), *The SAGE Handbook of Social Network Analysis*, 271–285. London: SAGE Publications Ltd.

Epilogue

John Nerbonne

Vaulting Ambition

Abstract Grammar studies have had overly ambitious goals. Computational linguistics, grammatical theory and corpus linguistics increasingly avoid the claims and perhaps even the goals of comprehensiveness. While parse accuracy (really parse and disambiguation accuracy) was a hotly contested field in the 1990s in computational linguistics, progress has stagnated in grammar-based work, even with models that include hundreds of thousands of independent variables. In grammatical theory, transformational generative grammar now limits its interest to "core" processes, while alternatives such as construction grammar seem to foreswear comprehensive studies, at least implicitly. Corpus linguistics, the focus of this volume, has always been more modest, and while it draws on ever more impressive amounts of data (> 10^{10} words/tokens), it also includes a lot of work on grammar differences — a fascinating, but different subject — rather than what constitutes grammar. I'll argue here nonetheless that corpus linguistics has a very valuable additional task in verifying judgments of unacceptability.

Keywords Grammatical theory, computational linguistics, corpus linguistics, generative linguistics

1 Introduction: Where's grammar?

There are some confusing differences among the various research communities trying to understand grammar – both those focused on the purely scientific study of grammar, such as grammatical theory, cognitive linguistics, and corpus linguistics but also those whose interest is less direct, namely computational linguistics. The differences concern goals, but also methods and perspectives. My own perspective is likely to be dominated by computation, which is why I'll begin with computational linguistics, but I have also made modest contributions to grammar theory and to corpus linguistics, and I find work in all these traditions valuable and interesting. In spite of this general appreciation of a lot of the work I see, I'm also critical of several aspects of the subfields, especially about what's missing, and impatient about the fairly poor level of interaction among the communities.

Clearly, since computational linguists emphasize a processing or methodological point of view, and corpus linguists a data-oriented one, the pride of place in the discussion might seem to be due to grammatical theory. This paper will nonetheless first present the perspective from computational linguistics because I understand it best and because it provides a way to understand the other two a bit better, particularly the limitations on observation and accuracy.

It would be wonderful to close with a sketch of a perspective that might overcome some of these difficulties in and among the different research traditions. Wonderful, but unrealistic. Since this is a volume on grammar and corpora, I'll focus my – hopefully constructive – criticisms on corpus linguistics.

2 Computational linguistics

There is a good deal of work on syntax in computational linguistics, and some of it adopts the generative paradigm, so that it overlaps strongly with grammatical theory. I'll discuss that work in Sec. 3.2 below so that I can concentrate on some computational linguistics insights that might be better appreciated in the other fields.

There was a close connection between grammatical theory and computational linguistics for a long time, fueled by the goal of building a general-purpose language understanding system and informed by the view that inattention to grammatical distinctions would inevitably disrupt or blunt the process. Theoretical grammar was definitely seen as an authority on the sorts of distinctions that Flickinger et al. (1987), Nerbonne et al. (1993) and Oepen and Flickinger (1998) document in test suites for the purpose of evaluating grammar processing. Others had argued that one might best ignore grammar where speakers often also appeared to, for example the difference between explicitly including the complementizer *that* in sentences such as *Mary knew (that) Sue would leave*, but the late Ivan Sag was clever at showing how such seemingly inconsequential grammatical details could be crucial for interpretation (Flickinger et al. 1987: 4):

(1) Did Jones know the woman (that) was the project director?

Omitting the complementizer in the sentence above changes the meaning completely: with the complementizer in (1) we have a question about whether Jones is acquainted with someone who was the project director, and without it (1) is a question about whether Jones knows a certain fact, namely, that the woman was the project director. Ignoring grammatical details risked what Sag called "pernicious dysfunction".

It is important for our purposes to note that test suites were not taken from corpora, authentically occurring speech or text, but instead consisted of minimal

examples designed to determine whether grammatical processing systems were assigning correct analyses to sentences. It is therefore fair to say that they were inspired more by grammatical theory than by corpus linguistics.

Naturally, it was also understood that grammar constitutes only one module of a complicated system, which also needs to include a lexicon, a parser, semantic and pragmatic interpretation, and an interface to an application, but grammar and parsing were central in research, and the lines to non-computational grammar research were kept close.

2.1 Massive ambiguity

The statistical revolution in computational linguistics has changed this enormously, and not only for practical reasons, as is sometimes assumed, since the triggering insight chronologically followed the wish to explore practical applications. Computational linguists began to consider parsing naturally occurring text – originally, mostly newspaper text – intrigued by application possibilities but also by the scientific challenge of looking beyond what grammar theory had concentrated on. One crucial insight that emerged as naturally occurring data became a focus is that the degree of ambiguity one encounters rises sharply. In fact, the number of analyses assigned by a linguistically well-informed grammar rises exponentially in sentence length. Gertjan van Noord's ALPINO parser and grammar (essentially of a head-driven phrase structure grammar sort (Müller 2016)) for Dutch have been developed over a period of over twenty years, and he and his colleagues and students have been at pains to develop it in a linguistically responsible way, i.e., avoiding Sag's problem as much as possible (Van der Beek, Bouma and Van Noord 2002). ALPINO was originally developed to produce analysis trees (or labeled bracketings), but it was later re-engineered to produce dependency graphs à la dependency grammar (now the common basis of comparison in parsing). As Figure 1 shows, sentences up to about ten words long are not very ambiguous at all in ALPINO, but then things become confusing quickly. This is a common result in grammar-based processing.

A natural reaction of non-computational linguists to this massive ambiguity is often polite skepticism, much like the reaction of non-linguists to linguists' observations of ambiguity. While it would not be feasible to examine all of the analyses assigned by a linguistically informed parser, one can examine many of the options exemplified in the set of analyses and ask whether the options belong in a strict grammar. Abney (1996) does exactly this, examining a myriad of interpretations assigned to the following sentence:

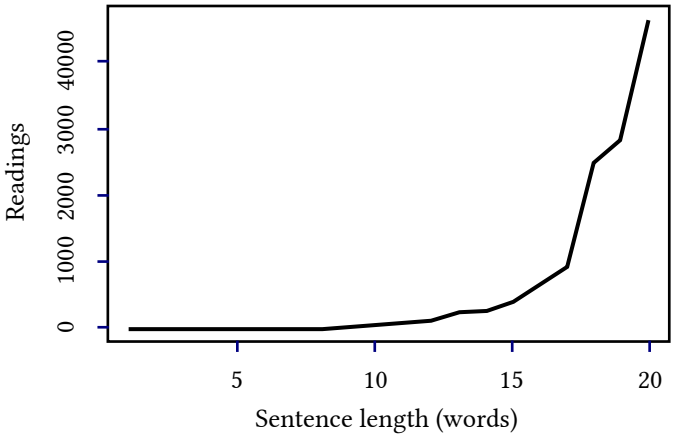


Figure 1: The average number of readings assigned by the linguistically well-informed grammar, ALPINO (Van Noord 2006). Note that 20-word long sentences (the length of an average newspaper sentence) have *on average* over 4,000 readings.

(2) In a general way such speculation is epistemologically relevant as suggesting how organisms maturing and evolving in the physical environment we know might conceivably end up discoursing of abstract objects as we do (Quine, Word and Object)

Abney points out that it is syntactically possible (albeit semantically nonsensical) to read the sentence in (2) so that *might* is a noun and *objects* a verb. *As we do* is naturally read as modifying the sentence or verb phrase headed by *discoursing*, but it could also modify the sentence or verb phrase headed by *objects*. The (computational) linguists that have tried to rule out ambiguities via stricter syntactic rules or selectional restrictions have generally given up, conceding that language is used flexibly enough to justify less strict rules.

The solution computational linguists turned to is STATISTICAL DISAMBIGUATION. In these systems, driven by machine learning (ML), one first collects a set of sentences and the analysis trees (or other analysis annotations) that correspond best to how each sentence is normally understood – one analysis per sentence with *no* indication of ambiguity. It would be unfeasible to ask human annotators to note the entire range of the thousands of analyses that sentences normally have, and the understood reading is most interesting in applications. The result is ANNOTATED DATA, which is used to train ML classifiers to choose which analysis tree best describes sentences that were not used in training. The annotated data is also used to evaluate how well sentences are parsed, and I will have more to say on this below. In this case, the data used for evaluation is withheld from the training data so as not to prejudice the evaluation (Black 1997). While the

annotations and therefore the testing were originally based directly on analysis trees (or, equivalently, on labeled brackets indicating analysis trees), the tree-like annotations in test material have largely given way to dependency labels (Briscoe et al. 2002). The points below do not hinge on the sort of annotation used.

From the point of view of grammatical theory, the move from categorical parsing to systems for parsing *cum* statistical disambiguation represented the loss, or at least the denigrating, of a valuable computational partner. The work of the 1980s involving categorical, non-statistical parsing was aimed at detecting errors in grammatical coverage in order to thereby improve grammars. The errors that were detected often led to discussion with theoretical grammarians about the sorts of distinctions and rules needed in the computational systems. However, the close collaboration was possible because test material consisted of minimal sentences designed to probe the discrimination of syntactic analysis systems. Once computational linguists began to work on real-world data (newspaper texts), the importance of length was brought home forcibly, as sentences in newspapers are about twenty words long on average. Pure grammatical analysis cannot be evaluated against all of the analysis trees produced by sentences working on newspaper text.

We note here that the original ambition of grammatically well-informed CL work had to be curbed – it turns out to be infeasible to check all the consequences of a grammar on all the data that is available.

2.2 Limited parsing accuracy

Computational linguists have agreed since the early 1990s that syntactic analysis systems needed to be evaluated strictly (Black, Lafferty and Roukos 1992). The discipline converged fairly quickly on a scheme borrowed from information retrieval in which both PRECISION and RECALL play a role.

In the precision-recall evaluation scheme, one parses a substantial amount of material – minimally a few hundred sentences, but often thousands – for which the correct analyses have been verified by humans. Let us focus for concreteness first on material that is annotated in labeled brackets. After the material has been parsed automatically by a system that is to be evaluated, one compares the results, constituent by constituent. A constituent the analysis assigns is regarded as correct in case the right label is assigned to the right sequence of words; anything else is incorrect. In particular we keep track of the size of the following sets:

1) the humanly annotated constituents the parser recognizes correctly (true positives, *tp*);

2) the humanly annotated constituents the parser failed to recognize (false negatives, *fn*);
3) the constituents postulated by the parser but not recognized by annotators (false positives, *fp*); and finally
4) the constituents not postulated by the parser and correctly not recognized by annotators (true negatives, *tn*).

Precision is then the fraction of analyses that are correct (recognized by human experts), *tp/(tp+fp)*, and recall is the fraction of the humanly recognized constituents that the parser detects, *tp/(tp+fn)*. The same sort of scheme may be applied to the currently more popular evaluation in terms of labeled dependencies (mentioned above), but I won't discuss this variant separately. We illustrate the analysis types with a tiny example (Figure 2, Table 1).

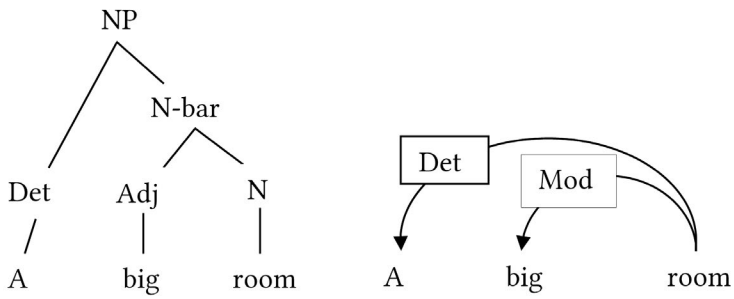


Figure 2: Syntactic analyses in test material against which parsers may be compared. Left, an analysis of constituency, and right, a (labeled) dependency graph. The tree on the left corresponds to the labeled bracketing [NP [Det a] [N-bar [Adj big] [N room]]], and the dependency graph shows a modifier dependency between the head 'room' and adjective 'big', and a determiner dependency between the same head and the determiner 'A'.

Table 1: The precision and recall rates of the analyses in the left column (parser output) based on the syntactic analysis above. I'm ignoring the non-branching nodes such as [Det a].

Parser Output	Prec.	Recall
`([Det a, big]), ([NP a,big,room])`	0.5	0.5
`([NP a,big,room])`	1.0	0.5
`([N-bar big, room]), ([NP a,big,room])`	1.0	1.0

Although I won't discuss separately the dependency-graphs as a basis for evaluation, I can mention that they're based on dependency triples of the sort "A-Det-room", and "big-Mod-room". One checks parser results against so-called "gold-standard" (human-annotated) structures, just as with the constituent-based evaluations.

There is a broad consensus in computational linguistics that these scores reflect parse accuracy faithfully, but those interested in grammar should keep concretely in mind what a (good) score of ninety percent (or *0.9*) means. A sentence *n* words long has *n-1* non-terminal (internal) nodes (if the nodes are binary branching, which is typical), so the average 20-word sentence from a newspaper corpus will have 17 constituents recognized correctly and two incorrectly. In other words, typical sentences will include some misanalysed nodes. (I'm ignoring the fact that errors don't appear uniformly, but instead tend to clump.) We return to this in Sec. 4 (below).

It is further worth mentioning that both precision and recall are measured because researchers may usually increase one at the expense of the other (Manning and Schütze 1999). Finally, there is an accepted way to combine the scores, F_1, a harmonic mean:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall}$$

Although parsing remains a central topic in computational linguistics, still attracting lots of energy and producing many papers (see the ACL Anthology[1]), the improvements in the last twenty years have not been substantial. F_1 rates for newspaper texts range from 0.89 to 0.92 for a variety of good systems (Ravi, Knight, and Soricut 2008), and no one expects rates to improve a great deal anytime soon.[2] Steedman (2011) argues that we're bound to see a decreasing rate of progress because improvements have occurred (linearly) as data reserves increased in size exponentially, and that further increases of the required size (an order of magnitude) are infeasible. Of course this is frustrating to those who'd hoped and aimed for grammar and parsing systems that assign exactly the right analyses to all the phrases and sentences in a language.

It may be interesting to corpus linguists to know that several researchers have experimented with the grammars implicit in corpora annotated for other purposes. In this sort of experiment, one extracts all the sub-trees of depth one, e.g. the two subtrees in Figure 2, *NP⟶ Det N-bar* and *N-bar⟶ Adj N,* together with their frequencies, and then uses these in a statistical parser. Klein and Manning (2003: 424) report an F_1-score of 0.726 using this approach. ML-based approaches are clearly doing a lot more than simply extracting rules and using their frequencies as estimates.[3]

1 https://aclweb.org/anthology/

2 Choe and Charniak (2016) have just published a paper enabling F_0 = 0.938 on a standard test set; this would be the best result and one of the biggest improvements in the last twenty years.

3 But see Charniak (1996) for an appraisal of "tree-bank grammars" more optimistic than Klein and Manning's.

2.3 Current work in computational linguistics

For experts in corpus linguistics and grammatical theory, it is worth knowing that substantial improvements have been made in domain adaptation, i.e. adapting a grammar and parser originally developed for newspaper text to domains such as Twitter, technical manuals or email (McClosky, Charniak and Johnson 2010), and in exploiting existing parsers to develop multilingual technologies. The latter effort is known as the "Universal Dependencies" project which proceeds from a cross-linguistically consistent treebank annotation for 70 languages and seeks to stimulate multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective (Nivre et al., 2016).

Finally, we cannot ignore in 2017 the ongoing shift in CL to the deep-learning methods inspired by neural networks (Socher et al. 2011, Schmidhuber 2015). These methods have already advanced to state-of-the-art practical use in some applications, e.g. machine translation (Wu et al. 2016), and they are contributing to modest improvements in parse accuracy, both in dependency parsing (Chen and Manning 2014) and in parsing categorial grammars (Ambati, Deoskar and Steedman 2016). However, improvements are difficult and modest (± 1 %). The developments are nonetheless worth following, since the results are being evaluated against the same sorts of annotated material used earlier. So improvements will accrue to tasks such as search for grammatical theory or corpus linguistics. A familiar complaint about neural-net-based processing would have it that the workings remain a black box, providing little insight as to why performance improves (when it does), but this criticism is being undercut by work showing how to analyze the inner workings of the networks a posteriori (Stoianov, Nerbonne, and Bouma 1998; Kuncoro et al. 2017).

This concludes our discussion of recent computational linguistics with regard to its relevance to the study of grammar. We return to generative models such as HPSG and LFG, which have a long history of collaboration with computational linguistics in Sec. 3.2 below.

3 Grammatical theory

In the first subsection below we consider work in the Chomskyan tradition, which I'll refer to as TRANSFORMATIONAL, in the second, work that strives toward more exactly formalized models (such as categorial grammar), and in the third work in cognitive linguistics.

3.1 The generative legacy

Grammar studies have inherited a well thought-out model from generative grammar including a statement of the task of grammar, namely to devise a completely explicit procedure capable of defining (or enumerating, or producing) all and only the grammatical structures in a language. Since sub-sentential structures occur within sentences, it is sufficient for the procedure to focus on producing – or GENERATING – sentences. And while the description of the task might sound at first pedestrian, even tedious, there is lots of room for theoretical discussion on what sorts of procedures, sub-procedures, modules, communication protocols, and even architectures are best suited for the task. If we date the beginning of Generative Grammar at the publication of Chomsky's *Syntactic Structures* (1957) then the model has inspired sixty years of research by thousands of researchers, and it has undoubtedly contributed enormously to the scientific understanding of grammar, its complexity, and to their implications for human cognition, learning and evolution.[4]

The operative goals of the enterprise have shifted over the decades, however, at least for most generative grammarians, both with respect to the range of phenomena considered and with respect to the level of detail of the description. Concerning the range of phenomena analyzed, attention focused increasingly on CORE GRAMMAR, essentially on constructions involving recursion. This was a conscious curtailment of the original descriptive ambitions of generative grammar (Pinker and Jackendoff 2005).

With respect to precision, early works emphasized the need for attention to concrete detail in generative theory building, in particular for specifying rules exactly and exhaustively. Most papers and even introductory texts from the last half of the generative period (since well before 1987, the mid-point between the publication of *Syntactic Structures* and now) are much less than explicit about the exact forms of rules, feature systems, and even grammar organization. The

4 See Roberta D'Alessandro's list "The achievements of generative syntax" for insights due to research in the transformational tradition. http://ling.auf.net/lingbuzz/003392 (viewed Apr. 11, 2017).

feeling of practitioners seems to be that more exact formulations can come later, that they may be abstracting usefully away from irrelevant details, and that the most pressing task currently is to understand the main lines of the overall system, e.g., the sorts of tree structures allowed (whether recursion is asymmetrically limited to the right edges of trees), whether tree transformations might be limited to a single rule or pair of rules ("merge"), or the nature of the information in the various modules and their interfaces, for which the concrete details in a broad coverage grammar are less than crucial. It also has turned out to be very difficult to obtain the abstract insights with any degree of certainty.

This wasn't always so. Early generative grammar was adamant about demanding exact and detailed formulations. For those who didn't witness this period personally, or who have forgotten it, the attitude was sharply different in the early days of this research line. When I worked on implementing grammar-processing systems in the 1980s and early 1990s, I kept a copy of Stockwell, Schachter and Partee (1973) on my desk for English, and Heidolph, Flämig and Motsch (1981) for German. The latter is not formalized, but the former is, and both present an enormous amount of very detailed material from the early generative tradition.[5]

Pullum (1989) complained about the decline in exact formulations, and documents the developments in the 1980s, stimulating an unusual response from Chomsky (1990), who basically defends the decreased level of formalization. Chomsky accepts the potential value of more formalization, but challenges that "the burden of proof is on those who consider the exercise worth undertaking." (p.146). It is interesting to note from our perspective that Chomsky (p.43) also speculates that the generative program as realized in his *The logical structure of linguistic theory* (1955) may have been "premature and far too ambitious".

Chomsky thus took issue with Pullum's complaint, but not with the observation that the methods have shifted in the generative grammar work that his ideas still dominated.[6] The main point for our argument is that detailed and explicit

5 *Mais l'honneur a ceux qui le meritent*! Broekhuis and Corver (2016) is the last of a seven-volume series on Dutch syntax from a transformational perspective and with an emphasis on description. See https://www.meertens.knaw.nl/cms/nl/medewerkers/143600-hansbr. It would be great to see more such works.

6 In the text I have refrained from stating that I agree with Pullum that there's a problem here. In fact, I agree that work from theoretical grammar I discuss in this section also suffers in quality because the link to the empirical basis of the claims is often vague. For the record, it's every researcher's right to research in the direction he or she finds most promising, most interesting, or that she judges she's most likely to be able to contribute to. Given Chomsky's enormous contributions, I would not presume to criticize his choice in how he conducts his research. The problem Pullum sketches arises not because of how a single researcher works but rather when no one

formulations are no longer being produced. Whether that's good or bad, it hinders the cooperation among computational linguists and corpus linguists, many of whom would like to profit from theory.

The shift from a focus on explicit and concrete detail has made the work less interesting to computational linguists working on syntax, virtually all of whom have turned to other research lines when they sought information from linguistics.[7] We turn to some of that work in the next section. The decision to shun concrete formulations has also made the generative work less interesting to all those who'd like to know the details of grammars and not just the main lines, which, as noted above, have also been difficult to pin down. Those interested in the details are not only computational linguists but also second-language instructors, language documentation specialists, language pathologists, and students of language contact and language change, all of whom work with more concrete details. The study of grammar would benefit if the channels of communication were more open.

3.2 Other generative traditions

Several competing frameworks have continued to insist on exact and detailed formulations, among them categorial grammar (CG), head-driven phrase structure grammar (HPSG), lexical-functional grammar (LFG), and tree-adjoining grammar (TAG), and in fact, many researchers in these frameworks rely on computational implementations of their research in order to test its coverage concretely, something which is virtually unknown in contemporary transformational work.

Stefan's Müller work on German grammar in the HPSG framework may serve as an example of the continued energy in this research line, but I hasten to add that there is excellent research in the other frameworks as well (Kaplan et al. 2004; Hockenmaier and Steedman 2002; Abeillé 1988; Kallmeyer and Osswald 2012). To begin, there are several extensive works on German syntax (Müller 1999, 2002, 2010) as well as a large number of detailed studies (see https://hpsg.hu-berlin.de/~stefan/Pub/ for more specialized studies), and Müller is at pains to compare these to work from the transformational community where he can (Müller 2016).

in a research community is taking advantage of the various modern means of quality control – corpus investigation (see below), strict formalization or computational implementation.

7 Eric Wehrli's work constitutes an honorable exception (Wehrli 1988). An assiduous referee pointed me to Abney and Cole (1985), Nelson (1987) and Kuhns (1986), all of whom implemented some aspects of government-binding theory, but as far as I know only Wehrli pursued this research line to the point of broad coverage.

Müller (2015) presents the computational system he uses to test his analyses, undoubtedly one of the reasons for the success his theoretical work has enjoyed. Carl Pollard once argued convincingly that formalization was a good way to detect inconsistencies and carelessness in theorizing, arguing the "PRO-theorem" (Chomsky 1986) was an example of how unformalized ideas could become confused. The "theorem" may be formulated very simply:

1. Every governed anaphor must be bound.
2. Every governed pronominal must be free (non-bound).
3. PRO is an anaphor and a pronominal.

PRO must not be governed.

The logic appears impeccable, but Pollard (1993) argues that it fails because the notion 'governed' is defined differently for anaphors and for pronominals. This means that the formulation effectively hides an equivocation. Rigorous formalization would promote the exposure of the slip, Pollard argues. I think that Pollard was right, but also that computational implementation is yet another, better tool in theory testing. Müller's work, like that of colleagues in categorial grammar, lexical-functional grammar and tree-adjoining grammars, generally exploits this tool.

The question nonetheless arises as to how these frameworks deal with the massive ambiguity and limited accuracy of modern natural language processing, and, of course they are limited in the same way as anyone else. Their systems, too, suffer from finding too many analyses for long sentences and are not more accurate when measured on free text. However, the parsing systems, when developed to support grammar research rather than for applications such as information extraction from newspaper text, can be tested on limited material especially designed for the topic at hand. Following the tradition of Montague Grammar, researchers generally implement FRAGMENTS of the grammar under study, not pretending to completeness (Müller and Lipenkova 2013). It is clear that this strategy risks inconsistencies across fragments, but the advantage is a more rigorous test of the (limited) grammar.

Corpora are properly the focus in Sec. 4 below, but the researchers pursuing alternative, stricter generative models have vigorously exploited corpora (Bildhauer 2011; Abeillé, Clément, and Toussenel 2003). The lines of communication are open between these communities.

3.3 Cognitive linguistics

Cognitive linguistics emphasizes that grammars are largely learned through experience and pleas therefore for a "usage-based" perspective (Ungerer and Schmid 2013), and it is especially popular among researchers in second-language learning (Robinson and Ellis 2008). Among grammatical theories, practitioners see the greatest affinity with Construction Grammar (Goldberg 2006), which rejects the old idea that fairly simple grammars rely on complex lexical information for treatments of compositional phrasal semantics and exceptionality (so-called "strict lexicalism").[8] Instead, each construction (or rule) may be associated with its own peculiar semantics and potentially idiosyncratic syntax. This perspective has stimulated hundreds of papers on various constructions, and has convinced most linguists that the older, strictly lexicalist view was flawed.

Construction grammar has also been stimulating theoretically. Stefanowitsch and Gries (2003) introduced collostructions into the grammarian's tool box, and Boyd and Goldberg (2011) show the need to tease apart the effects of sheer frequency in use (entrenchment) from the effects of encountering a form where another is expected (preemption). Preemption (also known as 'blocking') was well accepted in morphology, while the constructionists extended the applicability of the concept to syntax. There is also very interesting usage-based work on the effects of information density, but which doesn't identify Cognitive Linguistics or construction grammar as its inspiration. This is discussed in the following section.

In general, students of grammar appreciate not only detailed descriptions and analyses of individual constructions but also systematic treatments that attend to how constructions may interact and what sorts of constraints they might underlie. Berkeley Construction Grammar projected a systematic vision (Kay 2002), and Sag (2012) sketches a formalization within the HPSG feature formalism. While therefore more systematic developments of the Construction Grammar ideas exist, a great deal of the work focuses on the description of small numbers of constructions with no formal treatment.

If the usage-based perspective is maintained, one might easily imagine that "use" will turn out to break down into various sorts of uses, perhaps favoring different syntaxes, and that a closer cooperation with psycholinguists' work on sentence processing might develop (Pickering and van Gompel 2011), perhaps in particular in cooperation with models that involve memory crucially (Lewis and Vasishth 2005).

An encouraging development is Dunn's (2017) demonstration that construction grammars can be computationally learned, at least to some extent. While

8 See Müller and Wechsler (2014) for a dissenting view.

the result of the learning is not evaluated as parsers normally are (see above), Dunn does succeed in showing that instances of constructions are detected with measureable reliability. More work in this direction might accelerate collaboration among grammar theorists.

4 Corpus linguistics

Corpus linguistics has blossomed in the past thirty years, harvesting from ever larger corpora[9] and producing inter alia *the* standard grammar of English, built exclusively on corpus results (Huddleston and Pullum 2002). As noted above, there is an increasing amount of work in computational linguistics and in grammatical theory that tries to exploit corpus evidence extensively. The usage-based perspective of construction grammar lends itself immediately to corpus-based research, and this branch of theoretical linguistics has embraced corpora most enthusiastically (Gries and Stefanowitsch 2007), but, as we noted above, work in the alternative generative lines also exploits corpora where it can. Many researchers in grammar, particularly those who worked on languages they do not speak natively, welcome the opportunity to document that their non-native intuitions were not fantasies.[10]

Corpus linguists have also contributed theoretically, e.g. Biber and Conrad's (1999) ideas on lexical bundles, the affinities of words for other words, seen in recurrent sequences, even when they do not constitute idioms. Bresnan et al. (2007) demonstrate that a logistic regression model involving ten independent linguistic variables can predict the dative alternation with astounding accuracy (> 95%). The paper is a statistical tour de force, but we note that it is focused on explaining a variation in grammatical form rather than what constitutes grammaticality.

Another very interesting current research line has been enabled by work in corpus linguistics, even if it does not stem from it directly. Levy and Jaeger (2006), Jaeger (2010) and Linzen and Jaeger (2016) advance the thesis that texts tend to maintain a uniform information density, meaning that they distribute surprises (entropy peaks) fairly evenly. At least one large collection of research projects

9 As of Nov. 2016 DeReKo (IDS Mannheim) included over 30 billion word tokens.

10 An autobiographically inspired remark. My work on German impersonal constructions in the early 1980s met with skepticism about the acceptability of some data – even though it was taken in part from published works – because its author was not a native speaker. I was gratified when Hinrichs (2016) showed that the basic patterns are more widely attested. The general point is that corpora contribute in situations such as these as well.

is pursuing these ideas in concert ("Information Density and Linguistic Coding", http://www.sfb1102.uni-saarland.de/). Interestingly, Jaeger (2010: 24–25, 46–49) attributes some inspiration for his work to computational linguistics.

4.1 The status of corpus evidence

To avert misunderstandings, let's start with two points that might be platitudes, the first that corpus evidence is not completely reliable, just as evidence in general is not. Even well-edited newspapers and journals make mistakes, so that it always makes sense to examine crucial data and not simply assume that all that is published is well formed. Second, we acknowledge that the study of grammar cannot rely exclusively on corpus evidence. There are languages for which no corpora are available, and given the Zipfian distribution of words and constructions, rare structures, especially involving combinations of infrequent sub-structures, may simply not be instantiated, even in very large corpora. In addition, some of the methodology of grammatical theory goes well beyond that of checking whether or not a structure exists. I'm thinking of methods that ask whether one form is equivalent to another, whether one statement implies another, etc. We would impoverish grammatical theory if we made no use of that sort of data.

There are also numerous instances in which researchers have adduced corpus evidence that seems acceptable to the relevant judges and that contradicts putative generalizations. To someone who's worked with grammaticality intuitions, this does not seem surprising, as one often tries to construct data of a certain structure where it turns out that the judgments are less than robust. In particular, researchers working on languages other than their own are often leery lest they talk their respondents into preferred judgments or subliminally move them in a favored direction. So it's interesting to note that the more pressing problem is not obtaining an unbiased judgment of acceptability[11] but rather the failure to be imaginative enough to create well-formed examples of the sort under investigation. Intuition-based research is too quick to condemn. So Van der Beek et al. (2002) found examples of extraposition from comparatives in topic, and Meurers and Müller (2009) adduced violations of subjacency as well as doubly filled *Vorfeld* positions. The list may be extended easily, and I'd like to say more about this, given an attitude I've encountered among grammarians that

11 The experimental work showing that published grammaticality judgments tend to correlate with non-linguists' judgments (Schütze and Sprouse 2014, and references there) indeed ties intuitive data to a more general population, but note that it cannot address the problem that judges are too quick to condemn. That problem arises when researchers (or naïve judges) fail to consider a wide enough range of possibilities.

corpora cannot provide the "negative evidence" that plays such a central role in Generative Grammar (including both transformational and alternative generative theories of the sort discussed above). While it's true that corpora cannot *provide* negative evidence, they can contribute to *verifying* it. Corpus evidence shows that introspection is "too quick to condemn."

Bresnan (2007) reports on an especially interesting example involving the dative alternation, a construction which has been the subject of dozens, if not hundreds, of studies. The issue in focus was which verbs allow the "alternate" formulation. After searching for examples in corpora, she noted that she occasionally found *more* examples of the supposedly unacceptable sort than of the supposedly preferable one. One example involved the verb *drag*:

(3) … while Sumomo dragged him a can of beer (in corpora)
I dragged John the box ('' from published acceptability rating)

I tend to agree with the star in the example, albeit without great conviction, but I also find the dative-shifted example acceptable. The example is especially interesting because the construction is so well studied, but also because it illustrates succinctly one of the problems with negative judgments of acceptability, namely that they are based on concrete examples but are used to justify conclusions of general restrictions. A conclusion about the ungrammaticality of a construction (or rule) based on a single example is always a hasty generalization. Effectively, the unacceptability star in the example above was used to conclude that *drag* allowed only the PP indirect object, but not the double NP construction, a structural restriction. This conclusion is hasty without more extensive sampling, in this case with various combinations of arguments and adjuncts. Grammarians have been guilty of hasty generalizations when it comes to considering the import of unacceptability data.

Adli, García and Kaufmann (2015) document cases where critics of corpus linguistics methods have gone so far as to claim that intuitions and corpus evidence belong to different realms, so that corpus evidence could have no bearing on claims made in transformational research lines. Since the studies I surveyed adduced examples from corpora that pass the intuition test (they seem well formed), the move to dismiss all such evidence is hasty, and, given the "too-quick-to-condemn" problem, would render intuition-based theories methodologically ill-equipped. Finally, if it turned out to be necessary to isolate some theories from corpus evidence, that would *per se* make those theories less comprehensive and less interesting.

4.2 Shortcomings of corpus work today

Let's begin by reviewing known issues that ought to be solved. First, given the size of modern corpora, it seldom makes sense only to report statistical significance and not effect size. Despite the good advice of Baroni and Evert (2008), this still happens.

Second, search mechanisms are still not expressive (flexible) enough. Using XPATH and XQUERY, one can quantify over annotations, and this is useful if one wishes to study, e.g., impersonal constructions. Then one can search for nodes dominating finite verbs but which do not dominate subject complements (in the usual sorts of annotation). The following XPATH expression searches for the Dutch auxiliary verb *worden* that does not have a sister node with the dependency relation 'su' (subject)":[12]

```
//node[@lemma=”worden” and not(../node[@rel=”su”])]
```

It still finds some clauses with indirect objects that can look like subjects,[13] but it gets the job done. The key is in the interpretation of 'not' as 'there is no' (dependent subject). In TigerSearch one could use *atomic* negation (König and Lezius 2000: 6), which in the example above would amount to searching for nodes with the lemma *worden* and a dependent that is not a subject – which of course isn't the same thing. A search like the latter returns clauses with subjects as long as there are other dependents that are not subjects, e.g. direct or indirect objects. The negation was interpreted atomically, meaning 'with a dependent which is not a subject'. TigerSearch was an excellent tool in its time, but the query languages have improved. The improvements ought to be adopted more widely.

Third, and moving beyond the border of solved problems, we still need search interfaces for non-programmers. The query-by-example tool GrETEL (Augustinus et al. 2013, see too http://gretel.ccl.kuleuven.be) seems to be on the right track, but there may not be a perfect tool with respect to this issue.

However, there are much less tractable issues, too. Given the long tails in linguistic frequency distributions, large corpora are indispensable, and given the large size of contemporary corpora, only automatic annotation is feasible, i.e., annotation produced by taggers and parsers, not human judges. As we noted in the section on computational linguistics above, however, the accuracy of parses seems to have hit a ceiling still under 95% *per constituent*. This means that

12 With special thanks to Gosse Bouma, local XPATH guru.

13 Such as *u wordt verzocht* 'you are asked', but note *mij wordt verzocht* ('I(dat.) am asked'. These can be eliminated if one adds: and `not(../node[@rel=”obj2”])`).

sensitive work has to be checked manually for its dependence on potentially incorrect annotation. I don't suggest that everything needs to be checked, only samples, but I don't see any other way to immunize studies against annotation errors. It is worth adding here that parse errors and therefore annotations are not random noise, since some constructions, e.g., those involving coordination and ellipsis, are particularly error prone.

5 Conclusions and prospects

The direct communication between computational linguistics and grammatical theory is the focus of a relatively new journal, *Linguistic Theory and Language Technology*,[14] and this is surely a sign of serious interest. In particular, the reflective special issue on "The interaction of linguistics and computational linguistics" (Baldwin and Kordoni 2011) demonstrates that a number of prominent researchers in computational linguistics, including Ken Church, Eva Hajiová, Mark Johnson and Mark Steedman, advocate closer cooperation between the two fields, even if several authors see areas other than grammar as the most promising.

The transformationalists show less interest, and it is mostly critical. Everaert et al. (2015) criticize that computational linguists should focus more on hierarchical rather than sequential structure, but they select engineering-inspired work, which has practical constraints, to illustrate their points, and they ignore computational work on inducing phrase structure models, even in the only area that they discuss: machine translation (p.731).[15] The "deep-learning" models discussed in Sec. 2.3 above explicitly aim at modelling non-sequential phenomena.

My goal in writing this paper was to foster a bit more understanding about and among the various communities studying grammar. As I said above, virtually all students of grammar are taking advantage of corpora, and I also argued that corpora might be the only remedy to the "too-quick-to-condemn" problem. I don't think any special effort is needed to keep corpora in a central role in the study of grammar.

Of course, I won't try to argue that the field has been so ambitious that the allusion to Macbeth in the title is warranted. His admission to "vaulting ambition" closes a detailed, painfully honest reflection on why the actions he planned

14 *The Journal of Language Modeling* likewise aims "to help bridge the gap between theoretical linguistics and natural language processing (NLP)." (Przepiórkowski 2012).

15 Pereira (2000) suggests that a great many of the transformationalists' criticisms of work in computational linguistics implicitly assumes that computational linguistics never got beyond the stage of working on n-gram models of words. Everaert et al.'s paper is not reassuring on that score.

might be criticized. The colleagues in various sub-fields have also chosen their goals and methods judiciously.

Acknowledgments

I'm grateful to the following colleagues for discussion about the material discussed here, but, as usual, inferences as to the degree of agreement would be risky: Gosse Bouma, Edoardo Cavirani, Rob van der Goot, Martin Everaert, Göz Kaufmann, Stefan Müller, Gertjan van Noord, Marc van Oostendorp, Gertjan Postma, and Geoff Pullum. One anonymous referee was unusually helpful, and I thank him or her greatly, albeit anonymously.

References

Abeillé, Anne. 1988. Parsing French with Tree-adjoining Grammar: Some linguistic accounts. *Proc. 12th International Conf. on Computational Linguistics (COLING),* 7–12.

Abeillé, Anne, Lionel Clément and François Toussenel. 2003. Building a treebank for French. In Anne Abeillé (ed.), *Treebanks. Building and using parsed corpora.* 165–187. Netherlands: Kluwer.

Abney, Steven. 1996. Statistical methods and linguistics. In Judith Klavans and Philip Resnik (eds.), *The balancing act: Combining symbolic and statistical approaches to language*, 1–26. Cambridge: MIT Press.

Abney, Steven and Jennifer Cole. 1985. A government-binding parser. *Proceedings of the 16th Annual Meeting of the North-Eastern Linguistic Society.*

Adli, Aria, Marco García García and Göz Kaufmann. 2015. System and usage: (Never) mind the gap. In Aria Adli, Marco García García and Göz Kaufmann (eds.), *Variation in language: System-and usage-based approaches,* 1–25. Berlin: De Gruyter.

Ambati, Bharat Ram, Tejaswini Deoskar and Mark Steedman. 2016. Shift-Reduce CCG parsing using neural network models. *Proc. Conf. North American Chap., Association for Computational Linguistics: Human Language Technologies*, 447–453. Shroudsburg: ACL.

Augustinus, Liesbet, Vincent Vandeghinste, Ineke Schuurman and Frank Van Eynde. 2013. Example-based treebank querying with GrETEL–now also for spoken Dutch. *Proc. 19th Nordic Conference of Computational Linguistics* (NODALIDA 2013), 423–428. Linköping: Linköping University Electronic Press.

Baldwin, Timothy and Valia Kordoni. 2011. The interaction between Linguistics and Computational Linguistics. *Linguistic Issues in Language Technology* 6(1): 1–6.

Baroni, Marco and Stefan Evert. 2008. Statistical methods for corpus exploitation. In Anke Lüdeling and and Merja Kytö (eds.), *Corpus linguistics. An international handbook,* 777–803. Berlin: Mouton de Gruyter.

Biber, Douglas and Susan Conrad. 1999. Lexical bundles in conversation and academic prose. In Hilde Hasselgård and Signe Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson* (= Language and Computers 26), 181–190. Amsterdam: Rodopi.

Bildhauer, Felix. 2011. Mehrfache Vorfeldbesetzung und Informationsstruktur. Eine Bestandsaufnahme. *Deutsche Sprache* 4: 362–379.

Black, Ezra. 1997. Evaluation of broad-coverage natural-language parsers. In Giovanni Battista Varile et. al (eds.), *Survey of the state of the art in human language technology*. Cambridge: CUP.

Black, Ezra, John Lafferty and Salim Roukos. 1992. Development and evaluation of a broad-coverage probabilistic grammar of English-language computer manuals. *Proc. 30th Annual Meeting, Association for Computational Linguistics*, 185–192. Shroudsburg: ACL.

Boyd, Jeremy and Adele Goldberg. 2011. Learning what not to say: The role of statistical preemption and categorization in a-adjective production. *Language* 87(1): 55–83.

Bresnan, Joan. 2007. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. In Sam Featherston and Wolfgang Sternefeld (eds.), *Roots: Linguistics in search of its evidential base,* 75–96. Berlin: Mouton de Gruyter.

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

Briscoe, Ted, John Carroll, Jonathan Graham and Ann Copestake. 2002. Relational evaluation schemes. *Proc. Beyond PARSEVAL Workshop, 3rd International Conf. on Language Resources and Evaluation* (LREC), 4–8.

Broekhuis, Hans and Norbert Corver. 2016. *Syntax of Dutch: Verbs and verb phrases. Volume 3.* Amsterdam: Amsterdam University Press.

Charniak, Eugene. 1996. Tree-bank grammars. In *Proceedings of AAAI/IAAI*, 1031–1036.

Chen, Danqi and Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. *Empirical Methods in Natural Language Processing*, 740–750. Shroudsburg: ACL.

Choe, Do Kook and Eugene Charniak. 2016. Parsing as language modeling. *Proc. 2016 Conference on Empirical Methods in Natural Language Processing*, 2331–2336.

Chomsky, Noam. 1986. *Knowledge of language: Its nature, origin, and use.* Portsmouth, NH: Greenwood Publishing Group.

Chomsky, Noam. 1990. Topic...comment. On formalization and formal linguistics *Natural Language & Linguistic Theory* 8(1): 143–147.

Correa, Nelson. 1987. An attribute-grammar implementation of Government-Binding theory. *Proc. 25th Ann. Meeting, Association for Computational Linguistics*, 45–51. Shroudsburg: ACL.

Dunn, Jonathan. 2017. Computational learning of construction grammars. *Language and Cognition* 9(2): 254–292. Online (3/2016), DOI: https://doi.org/10.1017/langcog.2016.7

Everaert, Martin, Marinus Huybregts, Noam Chomsky, Robert Berwick and Johann Bolhuis, 2015. Structures, not strings: Linguistics as part of the cognitive sciences. *Trends in cognitive sciences*, *19*(12): 729–743.

Flickinger, Dan, John Nerbonne, Ivan Sag and Tom Wasow. 1987. Toward evaluation of NLP systems. Paper distributed at the 1987 ACL Meeting, Stanford. Avail. at www.let.rug.nl/nerbonne/papers/Old-Scans/Toward-Eval-NLP-1987.pdf

Goldberg, Adele E. 2006. *Constructions at work: The nature of generalization in language.* Oxford: Oxford University Press.

Gries, Stefan Thomas and Anatol Stefanowitsch. 2007. *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis.* Berlin: Walter de Gruyter,

Heidolph, Karl Erich, Walter Flämig and Wolfgang Motsch. 1981. *Grundzüge einer deutschen Grammatik.* Berlin: Akademie-Verlag.

Hinrichs, Erhard. 2016. Impersonal passives in German: Some corpus evidence. In Martijn Wieling, Martin Kroon, Gertjan van Noord and Gosse Bouma (eds.), *From semantics to dialectometry: Festschrift in honor of John Nerbonne*, 149–158. Milton Keynes: College Publications.

Hockenmaier, Julia and Mark Steedman. 2002. Generative models for statistical parsing with combinatory categorial grammar. *Proc. 40th Ann. Meeting, Association for Computational Linguistics*, 335–342. Shroudsburg: ACL.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge grammar of the English language.* Cambridge: Cambridge University Press.

Jaeger, T. Florian. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology* 61(1): 23–62.

Kallmeyer, Laura and Rainer Osswald. 2012. An analysis of directed motion expressions with lexicalized tree-adjoining grammars and frame

semantics. *International Workshop on Logic, Language, Information, and Computation*, 34–55. Berlin: Springer.

Kaplan, Ronald M., Stefan Riezler, Tracy King, John Maxwell III, Alexander Vasserman and Richard Crouch. 2004. Speed and accuracy in shallow and deep stochastic parsing. Xerox Palo Alto Research Center (PARC).

Kay, Paul. 2002. An informal sketch of a formal architecture for construction grammar. *Grammars* 5(1): 1–19.

Klein, Dan and Chris Manning. 2003. Accurate unlexicalized parsing. In *Proc. 41st Ann. Meeting, Association for Computational Linguistics,* 423–430. Shroudsburg: ACL.

König, Esther, and Wolfgang Lezius. 2000. The TIGER language--A description language for syntax graphs. (see CiteSeer^{x}, http://citeseerx.ist.psu.edu).

Kuhns, Robert J. 1986. A PROLOG implementation of Government-Binding theory. *Proc. 11th Intrnatn'l Conf. on Computational Linguistics.* (COLING), 546–550.

Kuncoro, Adhiguna, Miguel Ballesteros, Lingpeng Kong, Chris Dyer, Graham Neubig and Noah A. Smith. 2017. What do recurrent neural network grammars learn about syntax?. *Proc. 15th Meeting, European Chapter, Association for Computational Linguistics*, 1249–1258. Shroudsburg: ACL.

Linzen, Tal and T. Florian Jaeger. 2016. Uncertainty and expectation in sentence processing: Evidence from subcategorization distributions. *Cognitive Science* 40(6): 1382–1411.

Lewis, Richard L. and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29(3): 375–419.

Levy, Roger and T. Florian Jaeger. 2007. Speakers optimize information density through syntactic reduction. *Advances in neural information processing systems* 19, 849–856.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge: MIT press.

McClosky, David, Eugene Charniak and Mark Johnson. 2010. Automatic domain adaptation for parsing. *Human Language Technologies: The 2010 Ann. Conf. North American Chap., Association for Computational Linguistics*, 28–36. Shroudsburg: ACL.

Meurers, Walt Detmar and Stefan Müller. 2009. Corpora and syntax. In Anke Lüdeling and and Merja Kytö (eds.), *Corpus linguistics. An international handbook,* 920–933. Berlin: Mouton de Gruyter.

Müller, Stefan. 1999. *Deutsche Syntax deklarativ: Head-driven phrase structure grammar* für das Deutsche. Vol. 394. Berlin: De Gruyter.

Müller, Stefan. 2002. *Complex predicates: Verbal complexes, resultative constructions, and particle verbs in German.* Vol. 13. Stanford: CSLI.

Müller, Stefan. 2010. *Grammatiktheorie.* Tübingen: Stauffenburg.

Müller, Stefan. 2015. The CoreGram project: Theoretical linguistics, theory development and verification. *Journal of Language Modelling* 3(1): 21–86.

Müller, Stefan. 2016. *Grammatical theory: From transformational grammar to constraint-based approaches.* Berlin: Language Science Press.

Müller, Stefan and Janna Lipenkova. 2013. Chingram: A TRALE implementation of an HPSG fragment of Mandarin Chinese. *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, 240–249.

Müller, Stefan and Stephen Wechsler. 2014. Lexical approaches to argument structure. *Theoretical Linguistics* 40(1–2): 1–76.

Nerbonne, John, Klaus Netter, Kader Diagne, Judith Klein and Ludwig Dickmann. 1993. A diagnostic tool for German syntax. *Machine Translation* 8(1–2): 85–107.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. *Proc. 3rd Intrnatn'l Conf. on Language Resources and Evaluation* (LREC), 1659–1666.

Oepen, Stephan and Daniel P. Flickinger. 1998. Towards systematic grammar profiling. Test suite technology 10 years after. *Computer Speech & Language* 12(4): 411–435.

Pereira, Fernando. 2000. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 358.1769: 1239–1253.

Pickering, Martin and Roger van Gompel. 2011. Sentence parsing. In Matthew Traxler, and Morton Ann Gernsbacher (eds.), *Handbook of psycholinguistics,* 375–409. London: Academic Press (Elsevier).

Pinker, Steven and Ray Jackendoff. 2005. The faculty of language: What's special about it? *Cognition* 95(2): 201–236.

Pollard, Carl. 1993. On formal grammars and empirical linguistics In Andreas Kathol and Michael Bernstein (eds.), *ESCOL '93: Proc. of the 10th Eastern States Conference on Linguistics.* Columbus: Ohio State University.

Przepiórkowski, Adam. 2012. Journal of Language Modelling. *Journal of Language Modelling* 1: 1–4.

Pullum, Geoffrey K. 1989. Topic... comment. Formal linguistics meets the boojum. *Natural Language & Linguistic Theory* 8(1): 137–143.

Ravi, Sujith, Kevin Knight and Radu Soricut. 2008. Automatic prediction of parser accuracy. *Proc. Conf. on Empirical Methods in Natural Language Processing* (EMNLP), 887–896. Shroudsburg: ACL.

Robinson, Peter and Nick C. Ellis (eds.). 2008. *Handbook of Cognitive Linguistics and second language acquisition.* Oxford: Routledge.

Sag, Ivan A. 2012. Sign-based construction grammar: An informal synopsis. In Hans Boas and Ivan Sag (eds.), *Sign-based construction grammar*, 69–202. Stanford: CSLI Publications.

Schmidhuber, Jürgen. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61: 85–117.

Schütze, Carson and Jon Sprouse. 2014. Judgment data. In Podesva, Robert, and Devyani Sharma, (eds.) *Research methods in linguistics,* 27–50. Cambridge: CUP.

Socher, Richard, Cliff Chiung-Yu Lin, Andrew Ng and Chris Manning. 2011. Parsing natural scenes and natural language with recursive neural networks. *Proc. 28th International Conf. on Machine Learning (ICML-11)*, 129–136.

Steedman, Mark. 2011. Romantics and revolutionaries. *Linguistic Issues in Language Technology*. 6(11):1–20.

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

Stoianov, Ivelin, John Nerbonne and Huub Bouma. 1998. Modelling the phonotactic structure of natural language words with simple recurrent networks. In Peter-Arno Coppen, Hans van Halteren and Lisanne Teunissen (eds.), *Computational Linguistics in the Netherlands 1997.* 77–95. Amsterdam: Rodopi. (= *Language and Computers: Studies in Practical Linguistics* 25).

Stockwell, Robert P., Paul Schachter and Barbara Hall Partee. 1973. *The major syntactic structures of English.* New York: Holt, Rinehart and Winston.

Ungerer, Friedrich and Hans-Jörg Schmid. 2013 ([1]1996). *An introduction to cognitive linguistics*. London: Pearson.

Van der Beek, Leonoor, Gosse Bouma and Gertjan van Noord. 2002. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde* 7(4): 353–374.

Van Noord, Gertjan. 2006. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin (eds), *TALNo6. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, 20–42.

Wehrli, Eric. 1988. Parsing with a GB-grammar. In Uwe Reyle and Christian Rohrer (eds.), *Natural language parsing and linguistic theories*, 177–201. Netherlands: Springer.

Wu, Yonghui, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

In recent years, the availability of large annotated corpora, together with a new interest in the empirical foundation and validation of linguistic theory and description, has sparked a surge of novel work using corpus methods to study the grammar of natural languages. This volume presents recent developments and advances, firstly, in corpus-oriented grammar research with a special focus on Germanic, Slavic, and Romance languages and, secondly, in corpus linguistic methodology as well as the application of corpus methods to grammar-related fields. The volume results from the sixth international conference *Grammar and Corpora* (GaC 2016), which took place at the Institute for the German Language (IDS) in Mannheim, Germany, in November 2016. The editors of this volume are researchers at the IDS and were the organisers of *Grammar and Corpora 2016*.

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

ISBN 978-3-946054-82-5