Published in: Fuß, Eric/Konopka, Marek/Trawiński, Beata/Waßner, Ulrich H. (eds.): Grammar and corpora 2016. - Heidelberg: Heidelberg University Publishing, 2018. Pp. 11-24

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.