Introduction: Comparative Approaches to Data and Methods in Corpus Linguistics

Julia Schlüter and Ole Schützler

Motivations: Where We Aimed to Go, and Why

It is little wonder that corpus analyses have been, are and in all likelihood will for quite some time remain the staple method in linguistics: They are compatible with different theoretical approaches, they appeal to researchers invested in historical and present-day languages alike, they benefit from the constant increase in electronic databases and the digitalization of older texts, they are additionally boosted by advances in technology and software and the increasingly widespread *R*-literacy and programming skills among young-career as well as advanced linguists, and they draw inspiration from the knowhow of vibrant adjacent disciplines such as computational linguistics as well as commercial services on the Internet. Thus, the last one or two decades have seen a “paradigm shift” (Gries 2013: 4) towards ever more sophisticated quantitative techniques.

The present volume is motivated by the fact that approaches have diversified to an extent that necessitates more explicit discussion of methods in their own right, which must be accompanied by a critical reflection of what they can and cannot achieve, and perhaps of how they can complement each other. Most publications (with the exception of some monographs) inevitably and understandably devote little space to a detailed reflection or critique of methodological solutions or to the comparison of alternative approaches; even less often is the underlying rationale of the adopted approach discussed in any depth. Observing complex analytical tools in action may highlight what those methods can accomplish, but the reader faces only the end product, does not partake in the process of weighing alternatives and making choices, and may thus find it hard to draw the correct conclusions for his or her own research. This book addresses this lack of discussion at a relatively advanced technical level, at the same time explaining and questioning the more fundamental issues underlying the respective techniques. Our aim is to encourage a look beyond one’s own methodological horizon, to inspire critical understanding of and reflection on the work of others, to enlarge one's methodological repertoire and to enable informed choices between the available tools and methods.

Since making choices presupposes knowing the relevant differences between alternatives, we designed the book to follow a fundamentally comparative approach. This has been challenging for contributors (who deserve our sincere gratitude) as well as editors (more on which later), but – we believe – thoroughly beneficial.

The didactic orientation of the book is geared towards readers from the advanced graduate and doctoral levels upwards. It is of interest to scholars involved in any branch of corpus linguistics, synchronic variation and diachronic change (not only in English), and quantitative linguistics more generally. Elemental features of each chapter are the following:

- the inclusion of case studies, highlighting the interdependence of methodological and linguistic considerations
- the application of (at least) two methodological approaches to a single dataset or problem, or the exploration of (at least) two datasets for the same research question
- the explicit discussion and systematic comparison of strengths, weaknesses or simply different information values of the approaches
- the distillation of major conclusions in visually accessible synopses
- the inclusion of recommendations for further reading.

Several chapters take a fresh approach to issues that have long been central to corpus linguistics (e.g. choosing a corpus appropriate to a research question, choosing a methodology appropriate to a corpus, sampling texts into a corpus, sub-sampling from concordances, choosing baseline values for normalization, or integrating corpus metadata). Others address issues that promise to become more important in the future (e.g. planning a multifactorial approach to a dataset, choosing between frequentist and Bayesian statistics, or applying methods from computational linguistics and machine learning).

Despite being theory-neutral, it is of course true that corpus methods have been instrumental in promoting variationist theory-building by supporting the view that "language variation is systematic and can be described using empirical, quantitative methods" (Biber & Reppen 2015: 1–2). From the different case studies it will become evident that corpus methods can foster systematic accounts of diverse areas of linguistic variation and change, such as fine-grained pronunciation differences, (micro-)register differences in part-of-speech frequencies, geographical variation in speech prosody, quantitative diachronic drifts in grammar, subtle distinctions in the reflexes of pragmatic markers across languages, and the multifactorially conditioned selection of semantically more or less equivalent variants (such as bare and *to*-infinitives, *be-* and *get*-passives, *'s*-, *of-* and N-N-genitives, various

Cambridge University Press & Assessment
978-1-108-49964-4 — Data and Methods in Corpus Linguistics
Edited by Ole Schützler , Julia Schlüter
Excerpt
More Information

concessive constructions, etc.). Contributions to the volume do not fail to address the affordances and theoretical merits of the methodologies shown for their specific linguistic areas: Methodological discussion would indeed remain vacuous in the absence of concrete applications to actual data and their inherent challenges and pitfalls. The organization of the volume is however firmly based on methodological principles and the lessons that can be learnt for a large variety of applications from its distinctive comparative design, implemented in each of its chapters.

Signposts: A Roadmap for the Book

The present volume is arranged into four main parts, focusing on (1) fundamental characteristics of corpora (such as content, size and the amount and quality of annotation and metadata), (2) basic methodological constraints and choices in the selection and preparation of data for analysis, (3) alternative approaches to multifactorial analyses, along with an evaluation of their respective contributions and proposals for their further improvement, and (4) a look at classification-based techniques used in machine learning and computational linguistics but as yet less common in corpus linguistics. This partition identifies and makes transparent broader classes of methodological challenges, and it allows for direct access to and selective reading of individual topics of interest. The sequencing of the chapters constituting the parts also follows an underlying rationale, which can be summarized as follows.

Part 1 Corpus Dimensions and the Viability of Methodological Approaches

The opening part addresses considerations involved in the inevitable trade-off between the reliability of small and tidy corpora with various additional annotation levels on the one hand and the appeal of very large digital text databases that give access to fewer dimensions of linguistically relevant annotation. While the phonological focus of both chapters in this part is coincidental, the issues are of general methodological concern: Large databases may allow for automatized retrieval and analysis, and high type and token frequencies may statistically overrun inadequacies and errors in the database. However, excessive numbers of observations preclude a close-up analysis. What is more, the availability and validity of metadata will often be limited in big data, thus hampering a statistical analysis that controls for or pays due attention to, for example, speaker-, text- or genre-specific preferences. The chapters show that such characteristics of the database significantly determine the feasibility of methodological approaches; conversely, methods need to be adapted to the corpus or database at hand. Somewhat reassuringly, it turns out that in both case studies, the analyses of small and large databases can usefully complement each other and work synergetically.

The chapter by Lukas Sönning and Julia Schlüter is based on two standard reference corpora, the British National Corpus and the Corpus of Contemporary American English as opposed to the multi-billion-word database of Google Books Ngrams, which has, despite its allure, not been used in many systematic linguistic studies so far. Focusing on indefinite article allomorphy (*a* vs *an*) as an orthographic cue to the phonological strength of [h]-onsets in British and American English, the size advantage of the Ngrams database expectedly plays out in larger type and token counts, more stable estimates and fewer distortions due to data sparsity. However, as metadata are extremely limited (to year and variety), a fully accountable analysis is precluded from the outset. The case study illustrates how richly annotated corpora can shed light on potential disturbances arising from two sources: genre differences and between-author variability. A sensitivity analysis offers some degree of reassurance when extending the analysis to the Ngrams database. This allows the authors to demonstrate that the strengths and limitations of corpora and big data resources can, with suitable precautions, be counterbalanced to answer questions of linguistic interest.

In their chapter, Sabine Arndt-Lappe and Sebastian Hoffmann compare approaches to studying the effects of the prosodic Principle of Rhythmic Alternation on the basis of two fundamentally different corpus formats. The first is 'written' only (i.e. it consists of orthographic transcriptions of speech, or of originally written data), while the other one provides access to the sound files of the spoken data, too. The authors' main argument is, once more, that the nature and size of the corpus determines, or at least constrains, the range of methods that can be applied – and as a corollary of this, the findings that can be gained. Due to the greater availability and accessibility of written and transcribed spoken corpus data, much of the evidence in prior research is rather abstract and comes from large corpora accessed via the orthographic route only. Exploiting the recently available sound files of parts of the spoken section of the original British National Corpus, the authors analyse the data from an auditory perspective. This direct and highly controlled approach partially converges with and adds to the findings of previous studies. Thus, both approaches can be shown to complement each other, resulting in a better overall understanding of the phenomenon at hand.

Part 2 Selection, Calibration and Preparation of Corpus Data

A frequently neglected but consequential area in corpus-based studies concerns data preparation and the various decisions it involves. Rather than representing the language proper, corpora are sampled from a theoretically infinite range of actual and potential utterances. Sub-categories of text samples implemented in corpora are often taken for granted and uncritically accepted as bases for comparative studies (e.g. between varieties or genres). To avoid bias or error in analyses relying on these pre-defined distinctions, the representativeness of the selection made during corpus compilation may need to be ascertained. Further, for quantitative comparisons, the choice of appropriate baselines deserves critical attention. Once again, the standard of comparison has to be calibrated to the structure under investigation, and the most convenient measure (e.g. normalized frequencies) is not necessarily the best for valid results. The chapters in this part also draw attention to the fact that rather than being bags of words, corpora have a hierarchical structure, divided not only into spoken and written modes, (sub-)registers, age groups etc., but also into texts by individual authors or speakers, which may have a degree of internal consistency. The chapters discuss ways of accounting for such interdependence of data points when determining the confidence of our statistical estimates. Access to metadata (such as text IDs) and the dispersion of features across texts can also be essential in designing efficient subsampling schemes. The questions in this part are thus ‘what goes into a corpus’, as well as ‘what goes into an analysis’.

In his chapter, Fabian Vetter applies two methods to detect differences between corpus (sub-)registers, exemplified by the press editorials sections in the British, Canadian and Jamaican components of the International Corpus of English. By design, these methods are apt to target differences between varieties that are represented by putatively comparable corpus material, but it turns out that many of the observed differences can in fact be laid at the door of different sampling strategies applied by corpus compilers. In the example at hand, contrasts can be traced back to the division into institutional and personal editorials. This finding gives rise to a call for a higher granularity of sampling schemes, richer metadata (e.g. on the situational characteristics of the language samples included), and better documentation. As for the methods chosen, Vetter demonstrates that corpus-driven profiling based either on POS monograms or on higher-level multi-dimensional analysis performs reasonably well, with smaller differences in robustness and computational expense.

Sean Wallis and Seth Mehl review different baselines for the study of alternant choices, emphasizing that normalization to a standard number of words – while straightforward in its application – will in many cases not provide a meaningful measure of frequency. Instead, the authors underline the need for a baseline indicating opportunities of use, such as phrase or sentence counts. Exemplifying their proposal with reference to *get*- and *be*-passives and the presence or absence of agentive *by*-phrases, they demonstrate a sequence of measures taken to make the quantities that are compared more meaningful and defensible, based on linguistically informed selections of baseline quantities (number of main verbs, passives or potentially alternating passives). Crucially, this process must involve a categorization of observations by the researcher to ensure that mutual substitution is plausible in each case. To calibrate this manual data verification exercise to a manageable level, the authors apply a method of uneven category subsampling to the data, and use it to adjust variance estimates and confidence intervals in their analysis.

Lukas Sönning and Manfred Krug join the call for rich corpus metadata, also voiced by Fabian Vetter. They throw into relief the importance of the link between corpus hits and their sources (i.e. texts or speakers) by comparing study designs uninformed and informed by such metadata. Their argument draws attention to the consequences of uneven distributions of observations across the basic text units of a corpus. In the case study (a distributional analysis of the use of *actually* in the BNCs of 1994 and 2014), it is demonstrated that disproportionate word counts contributed by individual sources (in this case, speakers) will distort estimates for relevant subsections of corpus data (in this case, demographic groups defined by age and gender) if the analysis assigns the same weight to every observation. The proposed solution is to factor in the text (or speaker) level, but this hinges on the availability of the relevant metadata. Moreover, insights into the hierarchical structure of corpus data are shown to benefit the design stage of a study. Thus, if manual post-processing steps preclude an exhaustive analysis of corpus hits, insights into the organization of data points can direct down-sampling strategies to generate a statistically efficient subset of tokens.

Part 3 Perspectives on Multifactorial Methods

Various advanced statistical methods can be used to assess the impact of factors that simultaneously affect variable linguistic contexts. Typically, a corpus-based study will adopt a single methodological approach, which entails a certain perspective on the data and determines and restricts possible interpretations. The choice should therefore be deliberate and informed by an in-depth understanding of available techniques, rather than implicit or a matter of convenience. What unites all contributions in this part is that they take multiple approaches to testing for factor effects, along with a critical evaluation of assets and limitations of the respective methodologies. The approaches exemplified vary strongly, depending on the nature and dimensionality of the outcome variable(s) to be investigated. The challenges encountered in the process as well as in the interpretation of the results are illustrated and compared, affording concrete insights into the fundamental practical and conceptual differences between alternative approaches. Pushing the analysis one step further, chapters also develop aggregative and evaluative perspectives on two of the most recognized analytic approaches, regression analysis and distance-based visualization. For the former, it is discussed how the analysis can be made more robust in the case of sparse data and be supported with previous evidence; for the latter, visual data inspection is supported by mathematical diagnostics.

In his comparison of generalised linear mixed-effects models, generalised linear mixed-effects model trees and random forests, Tobias Bernaisch applies the three methodologies to a binary variable from the field of interactional pragmatics, the choice between filled and unfilled pauses across varieties of English represented by components of the International Corpus of English. He annotates a large number of examples for linguistic and extralinguistic factors (e.g. word class, gender or speaker age) and demonstrates the steps and decisions involved in the analyses. Though different in essence, the three resulting models share central trends. A more fine-grained evaluation of results and interpretations shows, however, that the three approaches differ in their systematicity of handling multiple observations from the same source, in that only the mixed-effects models explicitly account for and systematically partial out the relatedness of data points contributed by the same speaker. As to the way the approaches balance researcher involvement and control of the outcome, the approaches also differ substantially. Bernaisch shows how a modelling choice can thus lead to notably different perspectives on an identical set of data and variables.

Shifting the discussion from the previous chapter to a more complex scenario, Matthew Fahy, Jesse Egbert, Benedikt Szmrecsanyi and Douglas Biber work with a similar set of methods, but apply it to a ternary outcome variable: the variation between the *'s*-genitive, the *of*-genitive and functionally equivalent noun + noun combinations. The statistical approaches discussed fall into regression models on the one hand and classification trees on the other. Specifically, as an alternative to successive binomial regression analyses, the authors implement a multinomial model, which can analyse the entire dataset with three outcome categories simultaneously. Further, a basic classification tree is calculated alongside a more complex (and more robust) random forest. The chapter does not only weigh advantages and shortcomings of all four models, but it also explicates the different rationales and interpretations that come with them. As a major insight, it emerges that the nature of the dataset, the analytic purpose and the statistical model are interdependent and condition each other in several non-trivial respects.

Natalia Levshina's chapter compares standard frequentist and more recent Bayesian approaches to logistic regression analyses. Starting out from a multifactorial case study of the verb *help* complemented by either the bare infinitive or the *to*-infinitive, the key components and the main conceptual differences of frequentist and Bayesian inference are discussed. Conceptually, the Bayesian rationale of directly testing hypotheses on the effects of multiple factors on an outcome variable is argued to be preferable and more sensitive than the conventional approach of testing null hypotheses. On the practical side, Bayesian statistics enable the researcher to recycle and integrate the results of previous analyses based on different datasets as informative priors, which can help improve and stabilize statistical modelling. Recourse to prior research can thus produce synergies and reduce data preparation expense. In cases of data sparsity, it can by the same token enable researchers to analyse small samples. Bayesian methods are thus put forward as powerful tools for overcoming the limitations of isolated corpus studies and for promoting synergies between data collected by individual researchers.

In his chapter, Ole Schützler sets out by discussing the way in which multidimensional techniques and visualizations have been used to analyse linguistic data. While, for instance, multidimensional scaling and unrooted phenograms (or NeighborNets) have primarily been designed for exploratory purposes, he argues that they are in fact regularly used to put linguistic assumptions or hypotheses to the test. Cluster goodness (in terms of internal coherence and external distance from other clusters) in such approaches is typically evaluated based on a two-dimensional visualization. The author compares the affordances and limitations of visual inspection with a quantitative set of metrics that directly relates to visual displays but adds a degree of precision not attained by the human eye. The empirical part of the paper applies both approaches to a study of concessive constructions in six varieties of English, based on spoken and written material from the International Corpus of English. The author suggests that the new metrics can be usefully applied to a variety of multidimensional techniques to endow them with a greater measure of objectivity.

Part 4 Applications of Classification-based Approaches

The final part takes the examination of data and methods to regions where corpus linguistics meets computational linguistics and machine learning, two overlapping disciplines that have potential for supplementing and advancing linguistic research. The content of linguistic corpora can be processed using different categories of annotation (e.g. raw word forms, lemmas, part-of-speech tags and automatically inferred syntactic dependencies) and different levels of granularity (e.g. *n*-grams of different lengths). Rather than testing linguistic hypotheses directly, input generated in this way can be used to classify linguistic structures (or entire texts) based on the frequency profiles of categories at the respective level. In the case of preconceived divisions (e.g. into earlier and later corpus texts), an exhaustive computational analysis can output a virtually complete list of skewed unit frequencies. Alternatively, if linguists are looking for unknown correspondences (e.g. translation equivalents between parallel corpora), such an analysis can turn up a list of potential reflexes of source text structures in target texts. Since innovative approaches like those presented in these contributions do not by themselves ensure interpretability, it is indispensable to evaluate results in a linguistically informed perspective. The procedures showcased in this final part are thus exemplary in that they take their methodological inspiration from computational linguistics and discuss applications to corpus-linguistic tasks.

Focusing on grammatical changes in Late Modern and Present-Day English, Gerold Schneider applies a corpus-driven method to texts from two frequently used corpora for diachronic research, the Representative Corpus of Historical English Registers (ARCHER) and the Corpus of Historical American English (COHA). He compares his findings to those returned by more conventional corpus-based methods, which can be characterized as hypothesis-driven. To this purpose, the study employs automated profiling of large feature sets, such as word- and POS-based mono-, bi- and trigrams, chunks, syntactic dependency labels and measures of constituent order and length. The derived feature profiles are combined in a supervised classification task with a given division of texts into earlier and later corpus subperiods to reveal patterns of over- and under-use. Structures that are profiled as over- or under-represented in the diachronic subsections are then browsed for grammatical changes that may have been missed by previous research. According to Schneider, an advantage of such approaches is that they are theory-neutral and may generate novel hypotheses for investigation. These may then serve as input to further corpus-based approaches.

In the final chapter, Volker Gast addresses a problem of contrastive pragmatics: How can we study correspondences between pragmatic markers in two languages if one language has a class of elements that the other language lacks? Specifically, the contribution deals with modal particles of German (*ja* and *doch*) and their reflexes in English translations. As there is no predetermined set of potential English correspondences, traditional distributional analyses are not feasible, and methods from Natural Language Processing are explored instead. Using 32 types of *n*-grams, differing in length and type of annotation, three classification tasks are carried out (using a Support Vector Machine), in order to identify cues in the English translations that reflect the presence (or absence) of a particle in the German original. The results show that lemma-unigrams and -bigrams are often most informative, in the sense that they lead to the highest accuracy scores, while trigrams and 1-skip-2-grams provide important information about concomitants of modal particles that unigrams and bigrams miss. The results show that linguistic observables (*n*-grams) as the basis of quantitative analyses need to be carefully selected and explored in terms of their contribution to linguistic analysis.

Online Supplements: For Those Who Want to Go the Extra Mile

Using the supplementary materials provided online (e.g. *R*-scripts, datasets and interactive tools), readers can (1) follow the steps described by the authors, which will further an understanding of the technical aspects of individual chapters; (2) use the same data as the authors but analyze them with different analytical tools, thus taking the comparative approach beyond what is presented here; or (3) apply the methods that are described to their own data, thus generating additional test cases. Individual chapters will contain direct links to their respective online resources. All can additionally be accessed from a central portal site located at www.cambridge.org/schuetzler-schlueter.

Apart from the immediate didactic benefit, sharing materials in the online appendices available for each chapter is intended as a response to the call for open science and the challenges of what some perceive as an incipient “replication crisis” in linguistics (cf. Winter 2020: 47; cf. also Wallis 2021: 195–204): Readers are invited to study, recycle and refine datasets for their own research purposes or even to engage in a constructive dialogue with the contributors.

Reflections: Looking Back, Looking Ahead

From the editors’ perspective, this volume has turned out to be even more demanding in terms of time and effort, but eventually also more rewarding than anticipated. The challenges lay primarily in the methodological focus itself. Firstly, established assumptions were subjected to critical study and evaluation, which involved thinking out of the box and shaking up some long-established foundations, both of which can be uncomfortable things to do. Secondly, the problems and methods under scrutiny were certainly highly diverse, despite the unifying, overarching goal of learning from a comparative approach. On the upside, this means – or so we would like to think – that we ended up with a selection that is representative of the wide range of approaches current in corpus linguistics today. On the downside, the sheer amount of conventions that are part and parcel of each approach, the half-obscure workings of preconfigured algorithms that we were trying to shed light on, the diversity of conceptualizations and the lack of predefined or universally applicable criteria certainly made it more difficult to ensure the unity of the volume and to render the description of methodological steps transparent and traceable. Thirdly, we clearly observed once again what was one of the points of departure for the entire project: We are wont to think of hypotheses and theories that concern linguistic phenomena, and we use empirical methods to support them; but turning the methodological toolkit itself into the object of investigation appears to be more of a challenge than we had expected. In many a contribution, the