Argumentum 22 (2026), 32–38
Debreceni Egyetemi Kiadó
DOI: 10.34103/ARGUMENTUM/2026/3

Open Access

Randi Reppen, Larissa Goulart & Douglas Biber (eds.): The Cambridge Handbook of English Corpus Linguistics

Cambridge: Cambridge University Press, 2026, 616 pp.

József Andor

University of Pécs, Department of English Linguistics
andor.jozsef2@gmail.com

More than ten years have passed since the publication of the first edition of the *Cambridge Handbook of English Corpus Linguistics*. The appearance of the second edition, amounting to 616 pages, is similar in size and format to that of the first volume (623 pages). Concerning the nature of its contents, the second edition contains the bulk of the types of chapters of the first edition, however, it discusses an updated version of their descriptive potential in depth, in addition to which a number of new topics have been added in fourteen new chapters, representing the nature of the complexity of the English language as used by its native and non-native speakers. The important issues involved concern the description and analysis of the essential features of layers of its variability and the degree of its representativeness in a totalistic manner. The 29 chapters of the volume have been placed into four Parts of unified topics, which concern the standard general issues of corpus-based linguistic investigations. Part I, (3 chapters), analyzes particular topics of the description of major phenomena of the lexical, phraseological, formal, and semantic as well as pragmatic representation of the language. Part II, including 9 chapters, and the 13 chapters of Part III, present a corpus-based investigation of the manifestation of linguistic varieties of English. Part IV, containing 4 chapters, describes English usage for various types of applicative purposes. As noted by the editors in the *Introduction*, the authors of the chapters have been asked to first critically review earlier approaches preceding the topic of their investigation, followed by providing a description and/or analysis of its empirically relevant case study, which organization carefully warrants the unity and balanced nature of description of each of the approaches represented in the volume.

The first chapter of Part I, by *Tony Berber Sardinha*, provides excellent argumentation concerning one of the major issues of corpus-based linguistic analyses, that of the requirement that it should provide the highest degree of representativeness of the linguistic phenomena investigated. The two major factors in discussing corpus representativeness, those of domain specificity and sample planning, are characterized with special reference to the role of operationalization as well as the levelling and structuring of the variable strata in the representation of textual subcomponents, emphasizing the essential, linguistically relevant features of the mode of distribution of textual phenomena, such as size, the balanced nature of their manifestation in types of categorial form, as well as the rate of the volume of their occurrence. In the case study presented, samples of conversational register taken from the 2014 version of the British National Corpus are analyzed, with an emphasis on the rate of occurrence, the factor of length, as well as the categorial distribution of the lexico-grammatical phenomena investigated.

In the second chapter by *Laurence Anthony*, quantitative vs. qualitative approaches to corpus-based data are analyzed via the application of online and offline tools and methods. Availability of both types of tools is listed, interpreting the utilization of their strengths and weaknesses as

well as types of their application in detail, with special emphasis on describing the role of collocations and concordances related to single vs. multi-word forms of corpus-based appearance, forms of patterning, and the clustering of linguistically coded, usage-based facets and factors of representation, and providing argumentation over the status and role of keywords. Concerning the issue of keywordhood, however, the standard, purely quantitatively-based approach related to analyzing the corpus-based frequency of occurrence advocated by most corpus linguists (based on the pioneering view of Michael Scott (1997)) is accepted, rather than the textologically more relevant, conceptually grounded frame-based semantic-pragmatic potential of such lexical items or lexical composites influencing the mentally-based mode and factors of discourse structuring and its interpretation (Andor 2008).

The following chapter by *Stefan Th. Gries* introduces the concepts and utilization of quantitative designs as well as statistical techniques of the corpus-based analysis of linguistic data and the structuring of their textual representation. The author notes the requirement of statistically investigating the contextually-based types of information, the typology of lexical representation, as well as the frequency and dispersion of phenomena of the linguistically-based makeup of textual studies. Accounting for the even versus uneven distribution of items of textual representation, with special emphasis on the rate of occurrence of keywords, is essentially important in attempting to provide empirically realistic grounds for such investigations on the part of the corpus analyst. This chapter is basically informative rather than experimentally based, lacking discussion of observations related to an actual sample of discourse analysis presented in the form of a case study.

In Part II, various types of lexico-grammatical as well as prosodic phenomena of English are described and functionally discussed, chapter by chapter. The first of these, co-authored by *Karin McClellan* and *Shelley Staples*, provides a corpus-based investigation of the prosodic features of intonation in L1 and L2 speech, focusing on the patterning of tune, intonation-based phrasing, pitch range, and parameters of the rate of articulation, reviewing previous studies related to the regional varieties of English. New trends of investigation cover longitudinal studies, cases of bilingual contextual analyses, interfacing with issues of pragmatics, and the extension of the scope of analyses to the investigation of extralinguistic domains and underinvestigated languages. In their case study, the authors provide a functionally-based analysis of L1 vs. L2 variability using textual samples taken from the Louvain International Database of Spoken English Interlanguage, studying their lexico-grammatical fluency, the rate of fluency in their production, and related extralinguistic factors of manifestation such as gender-related differences in their representation.

In Chapter 5, co-authors *Paul Baker* and *Jesse Egbert* provide a brief investigation of using keyword analysis in corpus-based linguistics. Surveying previous approaches, they point to quantitatively-based aspects concerning the status and definition of this notion, based on the frequency of occurrence of such lexis in discourse-based corpora. Concerning future investigations, they rely on Egbert and Biber's (2019) approach rejecting the vital role of frequency and supporting the constructs of distinctiveness and the generalizability of keywords investigating their textual dispersion in analyses of the nature of the collocates and clusters of lexical items with a keyword role attributed to them in the analysis of discourse representation. In their case study, the authors present a keyword-based analysis, comparing the lexical structuring of textual samples of news discourse.

Chapter 6, co-authored by *Bethany Gray*, *Chris Nuttall*, and *Thomas Elliott*, investigates features of the collocational representation of phraseological constructs. They note that the categorization of such multi-word lexical patterns has been greatly varied in the linguistics literature. Their corpus-based vs. corpus-driven characterization appeared fairly recently. Major findings of the mode and typology of their collocational patterning have been gained from Biber's studies

of the occurrence of lexical bundles in various types of registers.[1] Recently, describing the nature of phraseological variability in such forms of usage has become a major issue of research, including the investigation of the expression of stance, the mode of discourse organization, and the functional load of referential framing. In the final section of the chapter, a case study of the role of the patterning of phraseological constructs is presented based on their occurrence in texts of academic writing vs. American English conversational discourse.

In the relatively short Chapter 7, *Magali Paquot* studies facets of the lexical and phraseological complexity of learner language, particularly the representation of sophistication in a multidimensional framework. The author emphasizes the role of its operation in L2 discourse, observing its manifestation in the lexical makeup of textual production at the C2 level of users.

In Chapter 8, *Douglas Biber* provides a corpus-based description and interpretation of grammatical variability in types of register-related usage and its occurrence in dialects, as well as the lexical relatedness of grammatical features. The scope of analysis is varied: it includes investigations of registers, genres, style, and text types. Special emphasis is given to studying the features of grammatical patterning phenomena in texts, specifically, contrasting the rate of occurrence of the categorical distribution of types of lexical representation in academic prose vs. samples of informal conversation. Biber's pioneering investigations provide the grounds for performing variationist sociolinguistics, giving a corpus-based extension to William Labov's classic observations of discourse representation with a similar scope of interest (Labov 2001). In the case study, he performs a variationist investigation into the textual realization of forms of the genitive construction.

In the following chapter, co-authors *Martin Hilpert* and *Christian Mair* describe the mode of studying the variability of grammatical change, with particular emphasis on understanding the process of the development of grammatical constructionalization and the nature of its occurrence in types of textual representation. The grammatical phenomena investigated include the grammatical coding of verbal complementation, the facts and factors of the usage of modal auxiliaries, changes of verbal inflections and of constructions of argument structuring. In the case study, a morphological issue, factors of the development of the English derivational suffix *-ment* are described and functionally interpreted, relying on data gained from the *OED* used as a corpus.

In Chapter 10, *Alan Partington* and *Anna Marchi* provide a corpus-based analysis of discourse representation describing and analyzing the discursive strategy of delegitimization of opposing speaker vs. author stances in argumentation functioning as a political strategy in textual representation, with special attention to the mode of the realization of persuasion. They point to the textual role of expressing facework in the expression of interpersonal communication. Their case study presents a comparative analysis of this domain in newspaper articles sampled from different English dailies, mainly reviewing its lexical manifestation of the content of 'scaremongering'. The aim of the authors with this bias was to provide a textually relevant 'macro analysis' rather than presenting a mere 'micro analysis' of discourse structuring and informative makeup. In my view, further extending their approach including a cognitively and conceptually relevant frame-based analysis on cognitive grounds would have significantly increased the interpretation of textual content in terms of the approach presented.

In the following chapter, *Anne O'Keeffe* and *Brian Clancy* introduce the role of corpus pragmatics in providing empirically valid grounds for the analysis of textual interpretation. It is pointed out by the authors that in earlier research, a dominant role was devoted to sentence-level rather than to discourse-level investigations, at best examining the co-textual features of the patterning of linguistic items, such as collocations, collostruction, colligation, and semantic prosody. Concerning the textual makeup of discourse, the nature and mode of turntaking strategies were also observed by researchers. In later and present-day research, the nature of multimodal corpora

[1] See, for instance, Biber (2006), and Biber and Conrad (2009).

has started to be investigated corpus-linguistically, with special attention devoted to observing the manifestation of the role of discourse-pragmatic markers, speech act representation, multimodality, the textual representation of keywords, and the expression of stance including, for instance, pragmatic conditions of the occurrence of politeness. In the case studies, a form-to-function analysis of the discourse representation of a single lexical item is presented, and the discourse-pragmatic marker status of *there* is analyzed.

In the last unit of Part II, Chapter 12, authored by *Chris Fitzgerald*, *Dawn Knight*, and *Svenja Adolphs*, aspects of analyzing multimodal corpora are discussed. Three types of multimodal corpora are studied: collections of video-recorded interaction of English found in Lingua Franca-based conversations, texts that contain visual elements such as dialogue acts, topic segmentation, expression of emotional states, and gestures, as well as corpora of signed languages. The case study presents a corpus-based analysis of turns occurring in virtual meetings.

The units of Part III of the volume contain corpus analysis of varieties of English usage. Chapter 13, by *Susan Conrad*, describes the nature of register variability, with special attention devoted to studying the manifestation of situation-based categorial components of textual representation, relying in her approach on Biber and Conrad's (2019) understanding of the notion of register. In surveying the relevant literature, she analyzes the register-related conditions of grammatical phenomena of English, the role of the manifestation of stance, the nature of co-occurring multiple linguistic features of textual representation. In the case study, three registers of texts of civil engineering are textually documented and analyzed.

In *Margaret Wood* and *Larissa Goulart*'s paper presented as Chapter 14, a multidimensional approach (MDA) to investigating the variability of textual representation is introduced. For almost three decades, researchers of MDA have provided feasible investigation of a range of variable spoken and written discourse yielding empirically grounded evidence of the mode of textual representation. The bulk of the chapter surveys the most important approaches of this field of research, with emphasis on discussing the variables of investigation conditions of tagger selection, as well as usage of appropriate, statistical methods. In their case study, the authors demonstrate a three-way comparison of the communicative purposes and representation of corpus samples, with special emphasis on analyzing their style, content, and their informational density.

Chapter 15 by *Shelley Staples* demonstrates modes and corpus-based techniques of the investigation of spoken discourse, advocating the benefits of Biber et al.'s (2022) register-functional framework complementing a multidimensional analysis to describe the lexico-grammatical complexity of English. According to this framework, as pointed out, two distinct types of complexity are relevant in the description of spoken vs. written registers of representation. Distinctive features of spoken discourse are operative in the expression of formulaic language, stance, sets of discourse markers, and vagueness of representation. Facts and factors of fluency and prosody will have to undergo further, empirically-based research, with reference to dialectal variability in future studies, as noted by the author. The case study focusses on identifying the critically important features of medical interactional discourse, with reference to text samples taken from the American Nurse-Standardized Patient corpus collected and transcribed in 2012, pointing to the importance of the manifestation of lexico-grammatical modal expressions in this type of discourse.

In the next section, Chapter 16, co-authors *Isobelle Clarke* and *Mark McGlashan* investigate the widespread nature of online language use: the Web and social media, with special attention to discussing the mode of corpus development deriving from online sources, the factors of the co-occurrence of keywords in usage, and the occurrence of cases of lexical innovations in discourse. They stress the need for relying on the usage of large corpora in such investigations aiming to trace and identify characteristic patterns of textual organization in particular websites. Concerning the identification of lexical items structuring in patterns in their textual occurrence, they fail to investigate their frame-based nature in expressing their keyword status in this mode

of their representation. In their case study, they present a corpus-based investigation of the functional role of the co-occurrence of keywords in scientific textual samples.

Chapter 17 by *Ken Hyland* presents argumentation over the usage of corpora in investigating the linguistically characteristic features of the textual register of academic English. It is pointed out by the author that the amount of corpus-based research into the patterning of academic English has been affluent in the past few decades. The permeation of academic writing by social interaction and intersubjectivity is also stressed. Following detailed surveying of the relevant literature, the major topics of investigation are reviewed, including frequency in usage, the representation and influence of keyness, the nature of concordancing, and the mode of annotation. In the case study, the character of various relevant types of corpora is briefly discussed.

Chapter 18 by *Eric Friginal* and *Ashleigh Cox* investigates the issue of the pedagogical relevance of professional registers, English for specific purposes, from a corpus-linguistics point of view. A survey of the past literature reveals that the major topics accounted for include business and finance, tourism, aviation, healthcare, and law, with emphasis on investigating the representation of context-specific norms in these types of texts. The rate of occurrence of characteristic types of pragmatically and semantically loaded words as well as of specific syntactic constructions and phrasal expressions has so far been mainly studied in corpora of relevant topical makeup. Emphasis is laid on investigating the nature of representativeness of such sources. In the case study, the discourse features of texts of eight professional registers undergo comparative analysis.

In the following Chapter 19, *Jesse Egbert* and *Thomas R. Lee* investigate institutional discourse, in particular, the representation of legal discourse. Concerning previous research, samples of judicial discourse, amicus briefs, and law review articles, the language of statutes and the text of the Constitution have been analyzed. In the empirical study attached to the detailed overview of key studies, COHA was used as a corpus source, aiming to observe the rate of occurrence of 'income' and the meaning of lexical expressions related to its content in a semasiological theoretical framework.

Chapter 20, co-authored by *Merja Kytö* and *Erik Smitterberg*, addresses the issue of the diachronic approach to the concept of registers. Two major parameters are researched in this field: considerations of the concept of domain, and the mode of the distribution of phenomena in the corpora applied. Additionally, the nature of their homogeneity and the typology of textual representation are also investigated. Concerning analyses of speech-related registers, the authors rely on the Corpus of English Dialogues as a major source of research. Finally, they provide a case study discussing the nature and mode of the factor of colloquialization.

In Chapter 21, *Michaela Mahlberg* presents a corpus-based study of the style of literary texts with reference to the seminal work of Leech (2008). Literary texts are considered to constitute a distinct type of register and in their related research they are compared in style, makeup and form to other types of registers. It is noted by the author that they formally occur as expressions of imagination and fiction. In their analysis statistical characterizations have been prevalent in addition to investigations of their narrative progression, lexico-phrasal patterning, concordance searches, cases of usage of lexical novelty, and the role of keywordhood. Investigations with this type of bias, unfortunately, lack consideration of the textual cohesive as well as coherence creating role of conceptual frames. In the last part of the chapter, Mahlberg investigates the textual patterning of Dickens's fiction.

Chapter 22, by *Jack Grieve* and *Dana Roemling*, addresses the issue of the corpus-based discussion of the natural language usage of dialect variation. In reviewing past research, the authors point to the questionnaire- and interview-based investigation of this domain, concentrating on sociolinguistic issues, including historically relevant aspects of studying regional dialect variation. In recent research of this area, approaching this domain has become computationally grounded, concentrating on describing the mode of appearance and role of the structuring of

networks, aspects of geographical distribution, and the characteristics of the occurrence of lexical innovations. The final section of the paper presents a corpus-based case study related to observing the distribution of phenomena of usage in varieties of American English.

In Chapter 23, *Marianne Hundt* presents a distributive analysis of the major characteristic features of varieties of World Englishes, including investigation of typologically and sociolinguistically relevant variation. An important aspect of her approach is that she provides accounts of types of usage in terms of the theory of construction grammar. In the case study, she provides a statistical, functionally based description of factors acting on the alternation of past vs. present perfect temporal expressions in the usage of L1 and L2 speakers.

Chapter 24 by *Philip Durrant* provides a pedagogically oriented approach to investigating the usage factors of English at L1 as used at school and university-based representation. The main topics of earlier research, as pointed out by the author, in addition to investigation of the language of mainstream students, have been studying the character of those with language impairment, such as individuals on the autism spectrum. The results of these studies call attention to the role of text genre on the content-based representation of usage. The field of these studies has been researching the representation of written rather than of spoken language. The topic of the author's case study is investigation of the lexico-grammatical features of usage in the linguistic performance of L1 speakers.

The following chapter, Chapter 25 by *Gaëtanelle Gilquin* and *Sylviane Granger*. is dedicated to describing the character of L2 English learners' linguistic performance. A general feature of the availability of learner corpora, as noted by the authors, is that written types are significantly higher in quantity than spoken ones, and also that these dominantly rely on data gained from advanced learners. Their research is mostly related to studying the discourse-based occurrence of elements with a connecting role, discourse as well as stance markers, features of involvement, and more recently, the modes of grammatical representation. Approaches to their investigation are primarily descriptive, predictive, and pedagogical. The empirical study presented provides a functionally based, comparative analysis of two-word discourse markers in native vs. non-native speech.

The last major unit of the volume, Part IV, contains four chapters reflecting on a wider, miscellaneous range of research topics.

The topic of Chapter 26 is analyzing the corpus-based status of dictionaries and lexicography, authored by *Anna Frankenberg-Garcia*. In the first part of the chapter the author reviews the history of changes from citation-based labeling of the definition of lexical content to its corpus-based identification. Current approaches to corpus-based description significantly rely on data of the frequency of occurrence of the lexical items represented, based on observing their collocational potentials in concordances. It is predicted by the author that in future developments in lexicography reliance on large databases investigated via the involvement of tools of AI will be decisive.

Chapter 27 by *Alex Boulton*, *Nina Vyatkina*, and *Thomas Cobb* describes a pedagogically grounded approach to the classroom application of corpus-based analysis. In the first section of the paper, reviewing the relevant literature, it is stated that the main aspect of the utilization of corpora by teachers lies in deciding what to teach. The second section presents a set of empirically-based methods of the usage of corpora in actual teaching, lacking concrete data of investigation.

In Chapter 28, *Randi Reppen* and *Fanny Meunier* present an overview and analysis of the application of corpus linguistics for general pedagogical purposes, to some extent similar in topic to the scope of Chapter 27. The aim of the authors is to point to the importance of Susan Conrad's work in this field and to changes that occurred after the appearance of her seminally important paper in 2000.

The last chapter of the volume by *Silvia Bernardini* and *Adriano Ferraresi* addresses the issues of a relatively new field: corpus-based translation studies, investigating the extent and mode of differences between translated and non-translated texts. One of the major, and perhaps most interesting topics of such research is identification of prototypical aspects of translation, relying on studying the factors of establishing cohesion and coherence to textual representations. Analysis of translator style has been extensively investigated in the past few decades. In the case study, the authors investigate the nature of collocations in translated language.

The handbook is closed with a brief index of names and terms.

To close this review, let me stress that the studies published in this volume are of extensively high quality, with a high degree of precision in representing data and the methods of the analyses presented. As a consequence, it is an extremely difficult task for the reviewer to find any lack of topical representation to be noted, proposals to undergo modification, or gaps in the argumentation presented in the chapters. I would recommend reading and studying the ideas represented in this handbook to students of English, theoretical as well as applied linguists, especially to those researching the application of corpora in linguistic description and analysis.

References

Andor, J. (2008): On the role of frame-based knowledge in lexical representation. In: Favretti, R. R. (ed.): *Frames, Corpora, and Knowledge Representation*. Bologna: Bononia University Press, 53–78.

Biber, D. (2006) *University Language: A Corpus-based Study of Spoken and Written Registers.* Amsterdam: John Benjamins. https://doi.org/10.1075/scl.23

Biber, D. & Conrad, S. (2009): *Register, Genre, and Style,* 1st ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814358

Biber, D. & Conrad, S. (2019): *Register, Genre, and Style*, 2nd ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, D., Gray, B., Staples, S. & Egbert, J. (2022): *The Register-functional Approach to Grammatical Complexity: Theoretical Foundation, Descriptive Research Findings, and Applications*. New York: Routledge. https://doi.org/10.4324/9781003087991

Conrad, S. (2000): Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly*, 34, 548–560. https://doi.org/10.2307/3587743

Egbert, J. & Biber, D. (2019): Incorporating text dispersion into keyword analyses. *Corpora*, 14(1), 77–104. https://doi.org/10.3366/cor.2019.0162

Labov, W. (2001): *Principles of Linguistic Change: Social Factors*. Oxford: Blackwell.

Leech, G. (2008): *Language in Literature: Style and Foregrounding*. Harlow: Pearson Education.

Scott, M. (1997): PC analysis of key words – And key key words. *System*, 25(2), 233–245. https://doi.org/10.1016/S0346-251X(97)00011-0