Introduction

Douglas Biber and Randi Reppen

Corpus linguistics is a research approach that facilitates empirical investigations of language variation and use, resulting in research findings that have much greater generalizability and validity than would otherwise be feasible. Studies carried out under the umbrella of corpus linguistics share certain research goals and distinctive analytical characteristics:

- they are empirical, analyzing the actual patterns of use in natural texts;
- they are based on analysis of a large and principled collection of natural texts, known as a 'corpus'; the corpus is evaluated for the extent to which it represents a target domain of language use;
- they make extensive use of computers for analysis, employing both automatic and interactive techniques;
- they depend on both quantitative and qualitative analytical techniques.

(see Biber, Conrad, and Reppen 1998: 4)

Corpus linguistics differs from other "hyphenated" areas of inquiry, such as sociolinguistics or psycholinguistics, in that it is not a theoretical subdiscipline of linguistics. That is, the prefixed element in hyphenated subdisciplines identifies the theoretical domain of inquiry: "socio-linguistics" is the study of language in relation to social factors; "psycho-linguistics" is the study of linguistic behavior in relation to psychological processes. But no such relation holds for "corpus linguistics." Rather, the distinctive characteristic of corpus linguistics is the claim that it is possible to actually "represent" a domain of language use with a corpus of texts, and possible to empirically describe linguistic patterns of use through analysis of that corpus. Any research question relating to linguistic variation and use can be approached from this methodological perspective.

This view of corpus linguistics is not universally accepted. For example, Stubbs (1993: 23–24) argues that "a corpus is not merely a tool of linguistic analysis but an important concept in linguistic theory," and Teubert

(2005: 2) describes corpus linguistics as "a theoretical approach to the study of language." However, this is a minority view, with most scholars focusing on the methodological strengths of corpus linguistics rather than treating it as a theoretical subdiscipline.

At the same time, nearly all scholars working in this area would agree that corpus linguistics is more than merely a methodological approach, because the analytical innovations of this approach have enabled researchers to ask fundamentally different kinds of research questions, sometimes resulting in radically different perspectives on language variation and use from those taken in previous research. Corpus linguistic research offers strong support for the view that language variation is systematic and can be described using empirical, quantitative methods. Variation often involves complex patterns of use that involve interactions among several different linguistic parameters but, in the end, corpus analysis consistently demonstrates that these patterns are systematic. In addition, corpus analyses have documented the existence of linguistic constructs that are not recognized by current linguistic theories. Research of this type – referred to as a "corpus-driven" approach – identifies strong tendencies for words and grammatical constructions to pattern together in particular ways, while other theoretically possible combinations rarely occur.

A novice student of linguistics could be excused for believing that corpus linguistics evolved in the past few decades, as a reaction against the dominant practice of intuition-based linguistics in the 1960s and 1970s. Introductory linguistics textbooks tend to present linguistic analysis (especially syntactic analysis) as it has been practiced over the past fifty years, employing the analyst's intuitions rather than being based on empirical analysis of natural texts. Against that background, it would be easy for a student to imagine that corpus linguistics developed only in the 1980s and 1990s, responding to the need to base linguistic descriptions on empirical analyses of actual language use.

This view is far from accurate. In fact, it can be argued that intuition-based linguistics developed as a reaction to corpus-based linguistics. That is, the standard practice in linguistics up until the 1950s was to base language descriptions on analyses of collections of natural texts: pre-computer corpora. Dictionaries have long been based on empirical analysis of word use in natural sentences. For example, Samuel Johnson's *Dictionary of the English Language*, published in 1755, was based on *c.* 150,000 natural sentences recorded on slips of paper, to illustrate the natural usage of words. The *Oxford English Dictionary*, published in 1928, was based on *c.* 5,000,000 citations from natural texts (totaling around 50 million words), compiled by over 2,000 volunteers over a seventy-year period (see the discussion in Kennedy, 1998: 14–15). West's (1953) creation of the *General Service List* from a pre-electronic corpus of newspapers was one of the first empirical vocabulary studies not motivated by the goal of creating a dictionary.

Grammars were also sometimes based on empirical analyses of natural text corpora before 1960. A noteworthy example of this type is the work of C. C. Fries, who wrote two corpus-based grammars of American English. The first, published in 1940, had a focus on usage and social variation, based on a corpus of letters written to the government. The second is essentially a grammar of conversation: it was published in 1952, based on a 250,000-word corpus of telephone conversations. It includes authentic examples taken from the corpus, and discussion of grammatical features that are especially characteristic of conversation (e.g. the words *well, oh, now,* and *why* when they initiate a "response utterance unit") (Fries 1952: 101–102).

In the 1960s and 1970s, most research in linguistics shifted to intuition-based methods, based on the theoretical argument that language was a mental construct, and therefore empirical analyses of corpora were not relevant for describing language competence. However, even during this period, some linguists continued the tradition of empirical linguistic analysis. For example, in the early 1960s, Randolph Quirk began the Survey of English Usage, a pre-computer collection of 200 spoken and written texts (each around 5,000 words) that was subsequently used for descriptive grammars of English (e.g. Quirk *et al.* 1972).

In fact, modern (computer-based) corpus linguistics also began during this period. Thus, work on large electronic corpora began in the early 1960s, when Kučera and Francis (1967) compiled the Brown Corpus (a 1-million-word corpus of published AmE written texts). This was followed by a parallel corpus of BrE written texts: the Lancaster–Oslo/Bergen (LOB) Corpus, published in the 1970s.

During the 1970s and 1980s, functional linguists like Prince, Thompson, and Fox also continued the empirical descriptive tradition of the early twentieth century, using (non-computerized) collections of natural texts to study systematic differences in the functional use of linguistic variants. For example, Prince (1978) compares the discourse functions of WH-clefts and *it*-clefts in spoken and written texts; Fox (1987) studied variation in anaphoric structures in conversational (versus written) texts; Fox and Thompson (1990) studied variation in the realization of relative clauses in conversation; Thompson and Mulac (1991) analyzed factors influencing the retention versus omission of the complementizer *that* in conversation.

What began to change in the 1980s was the widespread availability of large electronic corpora, and the increasing availability of computational tools that facilitated the linguistic analysis of those corpora. As a result, it was not until the 1980s that major linguistic studies based on analyses of large electronic corpora began to appear. Thus, in 1982, Francis and Kučera provide a frequency analysis of the words and grammatical part-of-speech categories found in the Brown Corpus, followed in 1989 by a similar analysis of the LOB Corpus (Johansson and Hofland 1989). Book-length descriptive studies of linguistic features began to appear in this period

(e.g. Granger 1983 on passives; de Haan 1989 on nominal postmodifiers) as did the first multidimensional studies of register variation (e.g. Biber 1988). During this same period, English language learner dictionaries based on the analysis of large electronic corpora began to appear, such as the Collins *COBUILD English Language Dictionary* (1987) and the *Longman Dictionary of Contemporary English* (1987). Since that time, most descriptive studies of linguistic variation and use in English have been based on analysis of an electronic corpus, either a large standard corpus (such as the *British National Corpus*) or a smaller corpus designed for a specific study. Within applied linguistics, the subfields of English for Specific Purposes and English for Academic Purposes have been especially influenced by corpus research, so that nearly all articles published in these areas employ some kind of corpus analysis.

Goals of the handbook

Basically, any research question or application relating to language variation and/or use can be approached from a corpus-linguistic perspective. Our goals in the *Cambridge Handbook of English Corpus Linguistics* (*CHECL*) are to survey the breadth of these research questions and applications in relation to the linguistic study of English. As such, the handbook includes chapters dealing with a wide range of linguistic issues, including lexical variation, grammatical variation, historical change, the linguistic description of dialects and registers, and applications to language teaching and translation. In each case, chapters assess what we have learned from corpus-based investigations to date, and provide detailed case studies that illustrate how corpus analyses can be employed for empirical descriptions, documenting surprising patterns of language use that were often unanticipated previously.

The goals of the *CHECL* are to complement, but not duplicate, the coverage of existing textbooks and handbooks on corpus linguistics. There are many excellent textbooks in print, providing thorough introductions to the methods of corpus linguistics, surveys of available corpora, and general reviews of previous research. The *CHECL* differs from these textbooks with respect to both the target audience and goals: the handbook is written for practicing scholars and advanced students in the field, offering a critical discussion of the "state of the art," rather than an introductory overview of the field in general. As a result, the handbook includes relatively little discussion of topics that have been fully covered in existing textbooks, such as surveys of existing corpora, or methodological discussions of corpus construction and analysis. Instead, the *CHECL* focuses on a critical discussion of the linguistic findings that have resulted from corpus-based investigations: what have we learned about language variation and use from corpus-based research?

The most innovative aspects of the *CHECL* are its emphasis on critical discussion, its explicit evaluation of the state of the art in each research area, and the inclusion of an empirical case study in each chapter. Although each chapter includes a broad summary of previous research, the primary focus is on a more detailed description of the most important corpus-based studies in this area, with discussion of what those studies found and why they are especially important. Each chapter also includes critical discussion of the corpus-based methods that are typically employed for research in this area, as well as an explicit summary of the state of the art: what do we know as a result of corpus research in this area that we did not know previously? Finally, each chapter includes an empirical case study illustrating the corpus analysis methods and the types of research findings that are typical in this area of research.

Organization of the handbook

As noted above, any research question relating to language variation and use can be approached from a corpus-linguistic perspective. In our previous work, we have identified two major objectives of such research:

(1) To describe linguistic characteristics, such as vocabulary, lexical collocations, phraseological sequences, or grammatical features. These studies often attempt to account for variation in the use of related linguistic features (e.g. the choice between simple past tense versus present perfect aspect) or to document the discourse functions of a linguistic feature.
(2) To describe the overall characteristics of a variety: a register or dialect. These studies provide relatively comprehensive linguistic descriptions of a single variety or of the patterns of variation among a set of varieties.

We have structured the main body of *CHECL* around these two domains of inquiry: chapters dealing with "Corpus analysis of linguistic characteristics" in Part II and chapters dealing with "Corpus analysis of varieties" in Part III.

Part II is organized as a progression of the linguistic levels, beginning with corpus-based analyses of prosodic characteristics, moving on to chapters dealing with lexical characteristics (keywords, collocations, and phraseology), followed by chapters on grammatical features (descriptive grammar, grammatical variation, grammatical change, and the intersection of grammar and lexis), and finally concluding with chapters on the corpus-based study of discourse functions and pragmatics.

Part III, then, is organized in terms of the range of varieties that have been studied from a corpus perspective. This part begins with chapters on the corpus-based description of spoken English, written academic English, and patterns of variation (synchronic and diachronic) among a wider range of spoken and written registers. Those chapters are then followed by

chapters on the use of corpus analysis to document the linguistic characteristics of other types of varieties: literary styles, regional dialects, world Englishes, English as a lingua franca, and learner English.

Preceding these two central sections, the *CHECL* has a shorter section dealing with methodological issues. As noted above, methodological issues relating to corpus design and analysis have been dealt with at length in previous textbooks. In addition, each of the chapters in *CHECL* includes discussion of the specific methodological considerations relating to their area of inquiry. However, beyond those treatments, there is need for a more general discussion of the current state of the art concerning corpus design and analysis. The three chapters included in Part I provide this discussion, dealing with current issues relating to corpus design and composition, tools and methods for the linguistic analysis of corpora, and quantitative research designs and statistical methods used to describe the patterns of use across corpora.

Finally, the *CHECL* concludes with a major section on applications of corpus-based research. Corpus linguistics has had a major influence on such applications over the past two decades, so that it is now almost impossible to find a research journal in applied linguistics, language teaching, translation studies, or lexicography that does not regularly publish articles utilizing corpus research findings. Part IV of the handbook surveys these major areas of application, including classroom applications, the development of corpus-based pedagogical materials, vocabulary studies, and corpus applications in lexicography and translation.

Internal organization of chapters

To help ensure the coherence of the *CHECL*, we have asked all authors to follow the same general organization in their chapter. While this has not always been possible, most chapters employ the same general organization. In addition to ensuring a coherent treatment across chapters, our primary goal is to provide a more sophisticated level of critical discussion than in most previous books. To achieve this goal, each chapter is composed of two major parts: a critical discussion of previous research, and presentation of an empirical case study.

Regarding the first section (the discussion of previous research), each chapter attempts to include the following:

- a general but concise survey of previous published research, briefly identifying the research topics covered by each study
- a more detailed discussion of the most important studies in this area: identifying the research questions; describing their methods; summarizing the major findings; and discussing why the study is especially important

- a critical discussion of the methods that are typically employed for research in this area, illustrated with more detailed discussions of studies that model strong research practices as well as studies that are problematic
- a summary of the state of the art for research in this area: what do we know as a result of corpus research in this area that we did not know previously? What are the major research gaps that still need to be addressed?

Regarding the second section (the empirical case study), each chapter addresses the following:

- a clear identification of the research question(s)
- motivation of the research question: why is the study important?
- a relatively detailed and critical description of methods: what are the strengths and weaknesses of the approach? Does it directly address the research questions? etc.
- a summary of the major research findings: what do we know as a result of this study that we did not know previously?

Our overall goal in requiring this strict organization across chapters is to achieve a handbook that will be of high interest to both students (with clear identification of the important research issues and discussion of strong and weak research practices) and advanced researchers (who can engage in the critical evaluations of each subfield).

Summary

In summary, the *CHECL* differs in three major ways from previous textbooks and handbooks on corpus linguistics. First, it has much more of a linguistic focus rather than a focus on the mechanics of corpus creation and analysis. Thus, most chapters in the *CHECL* deal with domains of linguistic inquiry, surveying the linguistic findings that have been achieved through corpus research.

Second, although methodological issues are important in the *CHECL*, they are addressed in each content chapter, rather than in isolation as topics in themselves. Further, these issues are addressed in a critical manner, evaluating the extent to which corpus designs and analysis techniques are in fact suitable for the linguistic research questions that are being investigated.

And third, the *CHECL* offers a more critical perspective than in most previous books. That is, rather than simply cataloging the range of research studies in an area of research, each chapter selects the most important of those studies, and describes the methods and research findings from those studies. Further, each chapter summarizes the state of the art in this area, describing what we have actually learned from corpus research. And finally, methods of corpus design and analysis are evaluated

critically with respect to specific linguistic research studies, to discuss the extent to which specific empirical research methods are well suited to the research questions of interest.

In sum, our goals in the *CHECL* go beyond a simple catalog of existing corpora and research tools, and go beyond simply itemizing the range of previous publications in this area. Rather, we hope to summarize and evaluate what we have learned about language use and variation from previous corpus-based research, to identify and discuss the most important of those previous studies and research findings, and to discuss the methodologies that work best for such research.

Part I

Methodological considerations