English Corpus Linguistics

An introduction

CHARLES F. MEYER

CAMBRIDGE

English Corpus Linguistics
An Introduction

CHARLES F. MEYER
University of Massachusetts at Boston

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2002

Printed in the United Kingdom at the University Press, Cambridge

Typefaces Times New Roman 10/13 pt. and Formata *System* LaTeX 2_ε [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Meyer, Charles F.
English corpus linguistics / Charles F. Meyer.
p. cm. – (Studies in English language)
Includes bibliographical references and index.
ISBN 0 521 80879 0 (hardback) – ISBN 0 521 00490 X (paperback)
1. English language – Research – Data processing. 2. English language – Discourse analysis – Data processing. 3. Computational linguistics. I. Title. II. Series
PE1074.5 .M49 2002
420'.285 – dc21 2001052491

ISBN 0 521 80879 0 hardback
ISBN 0 521 00490 X paperback

English Corpus Linguistics
An Introduction

English Corpus Linguistics is a step-by-step guide to creating and analyzing linguistic corpora. It begins with a discussion of the role that corpus linguistics plays in linguistic theory, demonstrating that corpora have proven to be very useful resources for linguists who believe that their theories and descriptions of English should be based on real, rather than contrived, data. Charles F. Meyer goes on to describe how to plan the creation of a corpus, how to collect and computerize data for inclusion in a corpus, how to annotate the data that are collected, and how to conduct a corpus analysis of a completed corpus. The book concludes with an overview of the future challenges that corpus linguists face to make both the creation and analysis of corpora much easier undertakings than they currently are. Clearly organized and accessibly written, this book will appeal to students of linguistics and English language.

CHARLES F. MEYER is Professor of Applied Linguistics at the University of Massachusetts, Boston. He has published numerous books and articles on linguistics, including *Apposition in Contemporary English* (Cambridge, 1992), and *The Verb in Contemporary English*, co-edited with Bas Aarts (Cambridge, 1995). He is currently editor of the *Journal of English Linguistics* and former co-ordinator of the International Corpus of English (ICE).

STUDIES IN ENGLISH LANGUAGE

The aim of this series is to provide a framework for original work on the English language. All are based securely on empirical research, and represent theoretical and descriptive contributions to our knowledge of national varieties of English, both written and spoken. The series will cover a broad range of topics in English grammar, vocabulary, discourse, and pragmatics, and is aimed at an international readership.

Already published

Christian Mair
Infinitival complement clauses in English: a study of syntax in discourse

Charles F. Meyer
Apposition in contemporary English

Jan Firbas
Functional sentence perpective in written and spoken communication

Izchak M. Schlesinger
Cognitive space and linguistic case

Katie Wales
Personal pronouns in present-day English

Laura Wright
The development of standard English 1300–1800: theories, descriptions, conflicts

STUDIES IN ENGLISH LANGUAGE

English Corpus Linguistics
An Introduction

To Libby and Freddie

Preface

When someone is referred to as a "corpus linguist," it is tempting to think of this individual as studying language within a particular linguistic paradigm, corpus linguistics, on par with other paradigms within linguistics, such as sociolinguistics or psycholinguistics. However, if the types of linguistic analyses that corpus linguists conduct are examined, it becomes quite evident that corpus linguistics is more a way of doing linguistics, "a methodological basis for pursuing linguistic research" (Leech 1992: 105), than a separate paradigm within linguistics.

To understand why corpus linguistics is a methodology, it is first of all necessary to examine the main object of inquiry for the corpus linguist: the linguistic corpus. Most corpus linguists conduct their analyses giving little thought as to what a corpus actually is. But defining a corpus is a more interesting question than one would think. A recent posting on the "Corpora" list inquired about the availability of an online corpus of proverbs (Maniez 2000).[1] This message led to an extensive discussion of how a corpus should be defined. Could something as specific as a computerized collection of proverbs be considered a corpus, or would the body of texts from which the proverbs were taken be a corpus and the proverbs themselves the result of a corpus analysis of these texts?

The answer to this question depends crucially on how broadly one wishes to define a corpus. The Expert Advisory Group on Language Engineering Standards (EAGLES) defines a corpus quite generally, saying that it "can potentially contain any text type, including not only prose, newspapers, as well as poetry, drama, etc., but also word lists, dictionaries, etc." ("Corpus Encoding Standard": http://www.cs.vassar.edu/CES/CES1-0.html). According to this definition, a collection of proverbs would indeed constitute a corpus. However, most linguists doing corpus analyses would probably prefer a more restricted definition of "corpus," one that acknowledged the broad range of interests among individuals who use corpora in their research but that defined a corpus as something more than a collection of almost anything. For the purposes of this book, then, a corpus will be considered a collection of texts or parts of texts upon which some general linguistic analysis can be conducted. In other words, one does not create a corpus of proverbs to study proverbs, or a corpus of relative

[1] Appendix 1 contains further information on the various corpus resources discussed in this book: Internet discussion lists such as "Corpora" as well as all the corpora described in this and subsequent chapters.

clauses to study relative clauses. Instead, one creates a corpus which others can use to study proverbs or relative clauses.

If a corpus is defined as any collection of texts (or partial texts) used for purposes of general linguistic analysis, then corpus linguistics has been with us for some time. Otto Jespersen's multi-volume *A Modern English Grammar on Historical Principles* (1909–49) would not have been possible had it not been based on a corpus representing the canon of English literature: thousands of examples drawn from the works of authors such as Chaucer, Shakespeare, Swift, and Austin that Jespersen used to illustrate the various linguistic structures he discusses. In recent times, a corpus has come to be regarded as a body of text made available in computer-readable form for purposes of linguistic analysis. The first computer corpus ever created, the Brown Corpus, qualifies as a corpus because it contains a body of text – one million words of edited written American English – made available in an electronic format (the ICAME CD-ROM, 2nd edn.) that can be run on multiple computer platforms (Macintosh, DOS/Windows, and Unix-based computers).

Modern-day corpora are of various types. The Brown Corpus is a "balanced" corpus because it is divided into 2,000-word samples representing different types (or genres) of written English, including press reportage, editorials, government documents, technical writing, and fiction. The purpose of designing this corpus in this manner is to permit both the systematic study of individual genres of written English and a comparison of the genres. In contrast, the Penn Treebank is not a balanced corpus: instead of containing a range of different genres of English, it consists of a heterogeneous collection of texts (totalling approximately 4.9 million words) that includes a large selection of Dow Jones newswire stories, the entire Brown Corpus, the fiction of authors such as Mark Twain, and a collection of radio transcripts (Marcus, Santorini, and Marcinkiewicz 1993). In creating this corpus, there was no attempt to balance the genres but simply to make available in computer-readable form a sizable body of text for tagging and parsing.

The Brown Corpus and Penn Treebank differ so much in composition because they were created for very different uses. Balanced corpora like Brown are of most value to individuals whose interests are primarily linguistic and who want to use a corpus for purposes of linguistic description and analysis. For instance, Collins (1991a) is a corpus study of modal verbs expressing necessity and obligation (e.g. *must* meaning "necessity" in a sentence such as *You must do the work*). In one part of this study, Collins (1991a) compared the relative frequency of these modals in four genres of Australian English: press reportage, conversation, learned prose, and parliamentary debates. Collins (1991a: 152–3) selected these genres because past research has shown them to be linguistically quite different and therefore quite suitable for testing whether modals of necessity and obligation are better suited to some contexts than others. Not only did Collins (1991a) find this to be the case, but he was able to explain the varying frequency of the modals in the four genres he studied. The fewest instances of

these modals were in the press reportage genre, a genre that is "factual, [and] non-speculative" and that would therefore lack the communicative context that would motivate the use of modals such as *must* or *ought*. In contrast, the conversations that Collins (1991a) analyzed contained numerous modals of this type, since when individuals converse, they are constantly expressing necessity and obligation in their conversations with one another. To carry out studies such as this, the corpus linguist needs a balanced and carefully created corpus to ensure that comparisons across differing genres of English are valid.

In designing a corpus such as the Penn Treebank, however, size was a more important consideration than balance. This corpus was created so that linguists with more computationally based interests could conduct research in natural language processing (NLP), an area of study that involves the computational analysis of corpora often (though not exclusively) for purposes of modeling human behavior and cognition. Researchers in this area have done considerable work in developing taggers and parsers: programs that can take text and automatically determine the word class of each word in the text (noun, verb, adjective, etc.) and the syntactic structure of the text (phrase structures, clause types, sentence types, etc.). For these linguists, a large corpus (rather than a balanced grouping of genres) is necessary to provide sufficient data for "training" the tagger or parser to improve its accuracy.

Even though descriptive/theoretical linguists and computational linguists use corpora for very different purposes, they share a common belief: that it is important to base one's analysis of language on real data – actual instances of speech or writing – rather than on data that are contrived or "made-up." In this sense, then, corpus linguistics is not a separate paradigm of linguistics but rather a methodology. Collins (1991a) could very easily have based his discussion of modals on examples he constructed himself, a common practice in linguistics that grew out of the Chomskyan revolution of the 1950s and 1960s with its emphasis on introspection. However, Collins (1991a) felt that his analysis would be more complete and accurate if it were based on a body of real data. Likewise, the computational linguist attempting to develop a tagger or parser could tag or parse a series of artificially constructed sentences. But anyone attempting this kind of enterprise knows that a tagger or parser needs a huge collection of data to analyze if it is expected to achieve any kind of accuracy.

Further evidence that corpus linguistics is a methodology can be found by surveying the various types of corpora available and the types of linguistic analyses conducted on them. The CHILDES Corpus contains transcriptions of children speaking in various communicative situations and has been studied extensively by psycholinguists interested in child language acquisition (MacWhinney 2000). The Helsinki Corpus contains various types of written texts from earlier periods of English and has been used by historical linguists to study the evolution of English (Rissanen 1992). The COLT Corpus (the Bergen Corpus of London Teenage English) contains the speech of London teenagers

and has been analyzed by sociolinguists interested in studying the language of a particular age group (Stenström and Andersen 1996). In short, linguists of various persuasions use corpora in their research, and are united in their belief that one's linguistic analysis will benefit from the analysis of "real" language.

If corpus linguistics is viewed as a methodology – as a way of doing linguistic analysis – it becomes increasingly important that corpora are carefully created so that those analyzing them can be sure that the results of their analyses will be valid. If a corpus is haphazardly created, with little thought put into its composition, then any analysis based on the corpus will be severely compromised. This book seeks to help corpus linguists understand the process of corpus creation and analysis by describing what exactly is involved in creating a corpus and what one needs to do to analyze a corpus once it is created. If corpus linguists understand the methodological assumptions underlying both the creation and subsequent analysis of a corpus, not only will they be able to create better corpora but they will be better able to judge whether the corpora they choose to analyze are valid for the particular linguistic analysis they wish to conduct. Although much of the discussion is relevant to the creation and analysis of any kind of corpus in any language, this book pays special attention to these issues as they apply to English language corpora.

To describe the process of corpus creation and analysis, I have divided this book into chapters that focus on the relationship between empirical studies of language and general linguistic theory, the considerations involved in the planning and creation of a corpus, the kinds of linguistic annotation that can be added to corpora to facilitate their linguistic analysis, and the process involved in analyzing a corpus once it has been created. In chapter 1 ("Corpus analysis and linguistic theory"), I discuss the role that corpora play in descriptive linguistic analysis and explore a controversy in modern-day linguistics that has been simmering since the rise of generative grammar in the 1950s: the conflict between the descriptive linguist, who often uses a linguistic corpus to produce descriptions of linguistic constructions, and the theoretical linguist, who stereotypically sits in his or her office contriving the sentences upon which some new theoretical point about language will be based. In this chapter, I argue that the corpus linguist and generative grammarian are often engaged in complementary, not contradictory areas of study: while the goals of the corpus linguist and the generative grammarian are often different, there is an overlap between the two disciplines and, in many cases, the findings of the corpus linguist have much to offer to the theoretical linguist. To illustrate how corpus analysis can benefit linguistic theory and description, I provide a sample analysis of elliptical coordinations that I conducted, and then give an overview of some of the corpora currently available and the types of linguistic analyses that they permit.

After discussing the role of corpus analysis in linguistics, in chapter 2 ("Planning the construction of a corpus"), I describe the various factors that have to be considered before the actual compilation of a corpus is begun. I discuss such considerations as how the corpus compiler determines the size of

a corpus, the types of texts that should be included in it, the number of samples for each text type, and the length of each text sample. Once decisions such as these are made, the actual creation of the corpus can begin, and in chapter 3 ("Collecting and computerizing data"), I provide advice on how a corpus can be most efficiently created. I discuss how to collect texts for inclusion in a corpus (i.e. make recordings and locate suitable written material), keep accurate records of the texts collected, obtain permission for written and spoken texts, and encode the texts in electronic form (i.e. transcribe spoken texts and optically scan printed material).

After a corpus has been created, its future use and analysis will be greatly facilitated if certain kinds of information are added in the form of linguistic annotation, the topic of chapter 4 ("Annotating a corpus"). In this chapter, I describe three kinds of annotation, or markup, that can be inserted in corpora: "structural" markup, which provides descriptive information about the corpus, such as the boundaries of overlapping speech segments in spoken texts or font changes in written texts; "part-of-speech" markup, which is inserted by software that automatically assigns each word in a corpus a part-of-speech designation (e.g. proper noun, modal verb, preposition, etc.); and "grammatical" markup, which is inserted by software that actually "parses" a corpus, identifying structures larger than the word, such as prepositional phrases or subordinate clauses.

While chapters 2–4 focus on the creation of a corpus, chapter 5 ("Analyzing a corpus") describes the process of analyzing a corpus. In this chapter, I conduct an actual corpus analysis to illustrate the various methodological issues that must be considered in any corpus analysis. I discuss how corpus analysts can best determine whether the size of the corpus they plan to analyze is suitable for the analysis being conducted, how analyses can be reliably conducted on different corpora collected under different circumstances, what software is available for assisting in the analysis of corpora, and once the analysis is completed, how the results of the analysis can be subjected to statistical analysis. In the final chapter, chapter 6 ("Future prospects in corpus linguistics"), I discuss where corpus linguistics is headed as a discipline, given projected developments in technology and the cost (in money and effort) it takes to create a corpus.

Although the approach I take in this book is relevant to the interests of a range of different corpus linguists, my primary focus is on how balanced corpora can be created and analyzed for purposes of descriptive linguistics analysis. For this reason, some topics are treated in less detail than they would be by corpus linguists with other interests. For instance, while the discussion of tagging and parsing in chapter 4 refers to work in natural language processing done in this area, I do not treat the topic of parsing in as much detail as a computational linguist designing parsers would. Likewise, in the discussion of statistics in chapter 5, there are many more statistical tests than I discuss that could have been covered. But the audience for whom these and other chapters were intended – linguists interested in creating and analyzing corpora – have more limited

interests in these areas. As a consequence, the areas are discussed in less detail, and more attention is given to actual linguistic analyses of corpora.

There are many people without whose advice and support this book would not have been possible. I am very grateful to Bill Kretzschmar, who encouraged me to write this book and who has offered many helpful comments on many sections. Merja Kytö, series editor for Studies in English Language, read the entire manuscript and provided feedback that has improved the book immensely. Two anonymous readers for Cambridge University Press read several draft chapters and gave me numerous comments that both strengthened the draft chapters and offered suggestions for completing the additional chapters I needed to write. Andrew Winnard, senior acquisitions editor at Cambridge University Press, provided expert guidance in taking the book through the review process. Others have given me very useful comments on individual chapters: Bas Aarts (chapter 1), Eric Atwell (chapter 4), Gerald Nelson (chapter 4), Robert Sigley (chapter 5), and Atro Voutilainen (chapter 4). Finally, I owe an extreme debt of gratitude both to my wife, Elizabeth Fay, who offered constant support, love, and encouragement during the years I spent writing this book, and to my son, Frederick Meyer, who at age three doesn't fully understand what corpus linguistics is but who has tried to be patient when I retreated to my study to sneak a few minutes to write this book.

Contents

	Preface	*page* xi
1	Corpus analysis and linguistic theory	1
2	Planning the construction of a corpus	30
3	Collecting and computerizing data	55
4	Annotating a corpus	81
5	Analyzing a corpus	100
6	Future prospects in corpus linguistics	138
	Appendix 1 Corpus resources	142
	Appendix 2 Concordancing programs	151
	References	153
	Index	162

1 Corpus analysis and linguistic theory

When the first computer corpus, the Brown Corpus, was being created in the early 1960s, generative grammar dominated linguistics, and there was little tolerance for approaches to linguistic study that did not adhere to what generative grammarians deemed acceptable linguistic practice. As a consequence, even though the creators of the Brown Corpus, W. Nelson Francis and Henry Kučera, are now regarded as pioneers and visionaries in the corpus linguistics community, in the 1960s their efforts to create a machine-readable corpus of English were not warmly accepted by many members of the linguistic community. W. Nelson Francis (1992: 28) tells the story of a leading generative grammarian of the time characterizing the creation of the Brown Corpus as "a useless and foolhardy enterprise" because "the only legitimate source of grammatical knowledge" about a language was the intuitions of the native speaker, which could not be obtained from a corpus. Although some linguists still hold to this belief, linguists of all persuasions are now far more open to the idea of using linguistic corpora for both descriptive and theoretical studies of language. Moreover, the division and divisiveness that has characterized the relationship between the corpus linguist and the generative grammarian rests on a false assumption: that all corpus linguists are descriptivists, interested only in counting and categorizing constructions occurring in a corpus, and that all generative grammarians are theoreticians unconcerned with the data on which their theories are based. Many corpus linguists are actively engaged in issues of language theory, and many generative grammarians have shown an increasing concern for the data upon which their theories are based, even though data collection remains at best a marginal concern in modern generative theory.

To explain why corpus linguistics and generative grammar have had such an uneasy relationship, and to explore the role of corpus analysis in linguistic theory, this chapter first discusses the goals of generative grammar and the three types of adequacy (observational, descriptive, and explanatory) that Chomsky claims linguistic descriptions can meet. Investigating these three types of adequacy reveals the source of the conflict between the generative grammarian and the corpus linguist: while the generative grammarian strives for explanatory adequacy (the highest level of adequacy, according to Chomsky), the corpus linguist aims for descriptive adequacy (a lower level of adequacy), and it is arguable whether explanatory adequacy is even achievable through corpus analysis. However, even though generative grammarians and corpus linguists have

different goals, it is wrong to assume that the analysis of corpora has nothing to contribute to linguistic theory: corpora can be invaluable resources for testing out linguistic hypotheses based on more functionally based theories of grammar, i.e. theories of language more interested in exploring language as a tool of communication. And the diversity of text types in modern corpora makes such investigations quite possible, a point illustrated in the middle section of the chapter, where a functional analysis of coordination ellipsis is presented that is based on various genres of the Brown Corpus and the International Corpus of English. Although corpora are ideal for functionally based analyses of language, they have other uses as well, and the final section of the chapter provides a general survey of the types of linguistic analyses that corpora can help the linguist conduct and the corpora available to carry out these analyses.

1.1 Linguistic theory and description

Chomsky has stated in a number of sources that there are three levels of "adequacy" upon which grammatical descriptions and linguistic theories can be evaluated: *observational* adequacy, *descriptive* adequacy, and *explanatory* adequacy.

If a theory or description achieves observational adequacy, it is able to describe which sentences in a language are grammatically well formed. Such a description would note that in English while a sentence such as *He studied for the exam* is grammatical, a sentence such as **studied for the exam* is not. To achieve descriptive adequacy (a higher level of adequacy), the description or theory must not only describe whether individual sentences are well formed but in addition specify the abstract grammatical properties making the sentences well formed. Applied to the previous sentences, a description at this level would note that sentences in English require an explicit subject. Hence, **studied for the exam* is ungrammatical and *He studied for the exam* is grammatical. The highest level of adequacy is explanatory adequacy, which is achieved when the description or theory not only reaches descriptive adequacy but does so using abstract principles which can be applied beyond the language being considered and become a part of "Universal Grammar." At this level of adequacy, one would describe the inability of English to omit subject pronouns as a consequence of the fact that, unlike Spanish or Japanese, English is not a language which permits "pro-drop," i.e. the omission of a subject pronoun that is recoverable from the context or deducible from inflections on the verb marking the case, gender, or number of the subject.

Within Chomsky's theory of principles and parameters, pro-drop is a consequence of the "null-subject parameter" (Haegeman 1991: 17–20). This parameter is one of many which make up universal grammar, and as speakers acquire a language, the manner in which they set the parameters of universal grammar is determined by the norms of the language they are acquiring. Speakers acquiring

English would set the null-subject parameter to negative, since English does not permit pro-drop; speakers of Italian, on the other hand, would set the parameter to positive, since Italian permits pro-drop (Haegeman 1991: 18).

Because generative grammar has placed so much emphasis on universal grammar, explanatory adequacy has always been a high priority in generative grammar, often at the expense of descriptive adequacy: there has never been much emphasis in generative grammar in ensuring that the data upon which analyses are based are representative of the language being discussed, and with the notion of the ideal speaker/hearer firmly entrenched in generative grammar, there has been little concern for variation in a language, which traditionally has been given no consideration in the construction of generative theories of language. This trend has become especially evident in the most recent theory of generative grammar: minimalist theory.

In minimalist theory, a distinction is made between those elements of a language that are part of the "core" and those that are part of the "periphery." The core is comprised of "pure instantiations of UG" and the periphery "marked exceptions" that are a consequence of "historical accident, dialect mixture, personal idiosyncracies, and the like" (Chomsky 1995: 19–20). Because "variation is limited to nonsubstantive elements of the lexicon and general properties of lexical items" (Chomsky 1995: 170), those elements belonging to the periphery of a language are not considered in minimalist theory; only those elements that are part of the core are deemed relevant for purposes of theory construction. This idealized view of language is taken because the goal of minimalist theory is "a theory of the initial state," that is, a theory of what humans know about language "in advance of experience" (Chomsky 1995: 4) before they encounter the real world of the language they are acquiring and the complexity of structure that it will undoubtedly exhibit.

This complexity of structure, however, is precisely what the corpus linguist is interested in studying. Unlike generative grammarians, corpus linguists see complexity and variation as inherent in language, and in their discussions of language, they place a very high priority on descriptive adequacy, not explanatory adequacy. Consequently, corpus linguists are very skeptical of the highly abstract and decontextualized discussions of language promoted by generative grammarians, largely because such discussions are too far removed from actual language usage. Chafe (1994: 21) sums up the disillusionment that corpus linguists have with purely formalist approaches to language study, noting that they "exclude observations rather than . . . embrace ever more of them" and that they rely too heavily on "notational devices designed to account for only those aspects of reality that fall within their purview, ignoring the remaining richness which also cries out for understanding." The corpus linguist embraces complexity; the generative grammarian pushes it aside, seeking an ever more restrictive view of language.

Because the generative grammarian and corpus linguist have such very different views of what constitutes an adequate linguistic description, it is clear