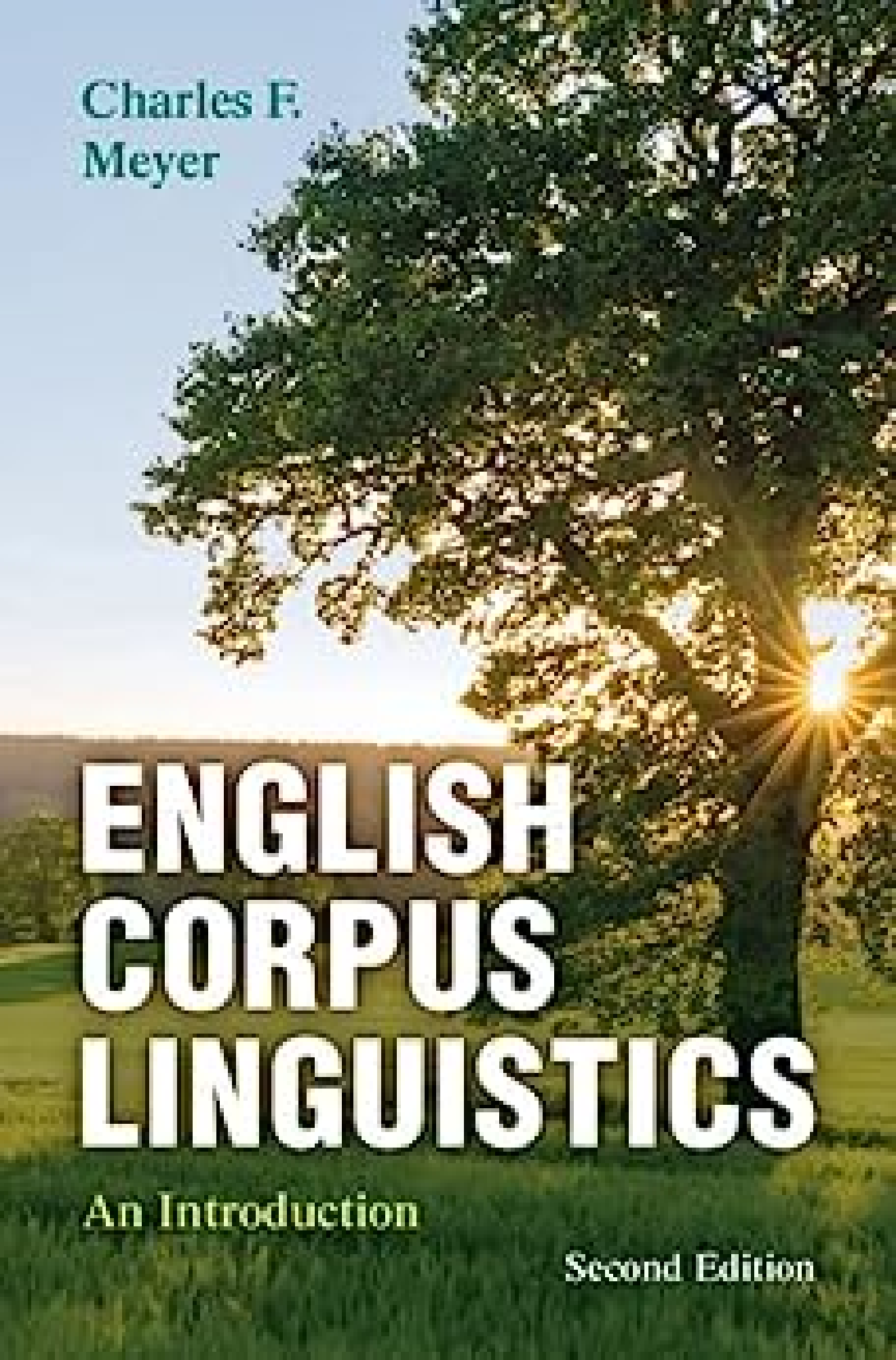
Charles F. Meyer
ENGLISH CORPUS LINGUISTICS
An Introduction
Second Edition

English Corpus Linguistics

Second Edition

Corpus linguistics is a research method which draws on authentic language examples, collected and organized into "corpora," or searchable "bodies" of data. The method was established in the 1960s and has rapidly developed since then. Now in its second edition, this book provides a step-by-step guide on how to create and analyze linguistic corpora. It has been extensively updated to reflect the most recent developments in this ever-evolving field, and now covers the empirical foundation of corpus-based research, new methodological considerations that guide the creation of a corpus, new kinds of research that can be conducted on corpora, and the most up-to-date information on how qualitative and quantitative analyses of corpora are conducted. Theoretical approaches are introduced in an accessible, easy-to-read way, and the book is illustrated with a wide range of different linguistic corpora, making it essential reading for researchers and students in a number of subfields of linguistics.

CHARLES F. MEYER is Professor Emeritus of Applied Linguistics at the University of Massachusetts, Boston. He has published extensively in the area of corpus linguistics and has been actively involved with the creation of corpora, such as the American component of the International Corpus of English.

English Corpus Linguistics
An Introduction

Second Edition

CHARLES F. MEYER
University of Massachusetts

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107057159
DOI: 10.1017/9781107298026

First published 2002
Second edition 2023

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Meyer, Charles F., author.
Title: English corpus linguistics : an introduction / Charles F. Meyer.
Description: Second edition. | New York, NY : Cambridge University Press, 2023. |
Includes bibliographical references and index.
Identifiers: LCCN 2022050194 (print) | LCCN 2022050195 (ebook) |
ISBN 9781107057159 (hardback) | ISBN 9781107681835 (paperback) |
ISBN 9781107298026 (epub)
Subjects: LCSH: English language–Research–Data processing. | English
language–Discourse analysis–Data processing. | Corpora (Linguistics) |
Computational linguistics.
Classification: LCC PE1074.5 .M49 2023 (print) | LCC PE1074.5 (ebook) |
DDC 420/.285–dc21/eng/20221123
LC record available at https://lccn.loc.gov/2022050194
LC ebook record available at https://lccn.loc.gov/2022050195

ISBN 978-1-107-05715-9 Hardback
ISBN 978-1-107-68183-5 Paperback

Contents

Figures

Tables

Preface

This book is an extensive revision of the first edition of *English Corpus Linguistics*. It reflects the many changes and developments that have taken place in the area of corpus-based research on the English language since the publication of the first edition in 2002. Not only are there currently many more corpora available for analysis but corpora have changed in other ways too.

Corpora have become larger. While the Brown and LOB (London-Oslo-Bergen) Corpora (released in 1962) were one million words in length, the Corpus of Contemporary American English (COCA) is currently one billion words in length. Moreover, COCA represents another change: the availability of corpora for analysis over the Web that are linked to web-based concordancing programs that can be used for analysis. In fact, the Web itself is now envisioned as a corpus. For instance, the WebCorp search engine (www.webcorp.org.uk/live/) can be used to search the Web for words or phrases. It can also be used to search specialized corpora taken from the Web, such as The Synchronic English Web Corpus (www.corpusfinder.ugent.be/synchronic-english-web-corpus), which contains 467,713,650 words from texts on the Web.

English-language corpora (and corpora in general) have become more diverse. The Brown and LOB corpora contained different genres of edited written American and British English, respectively. In the 1990s, the International Corpus of English (ICE) expanded the varieties of English included, containing samples of speech and writing from many international varieties of English, including British, US (written only), Indian, Singapore, Irish, and Hong Kong (cf. Greenbaum 1996a as well as www.ice-corpora.uzh.ch/en.html). The Corpus of Global Web-Based English is 1.9 billion words in length and consists of English taken from the Web representing 20 countries in which English is spoken. In general, language variation has become an important area of inquiry in the area of corpus-based research.

Linguistic diversity can be found in other types of corpora too. Parallel corpora contain two languages with one language translated into another language, and the two corpora aligned at the level of the sentence. Many such corpora contain English as one of the languages. For instance, the Europarl Corpus (Release V7) consists of transcriptions of 21 European languages taken from meetings of the European Parliament that were translated into English. Sentences are aligned so that English translations can be directly compared to the original sentences on which they are based. There are many learner corpora, which contain samples of English written by speakers for whom English is a second or foreign language. The ICLE Corpus (The International Corpus of Learner English) contains samples of written English produced by advanced learners of English as a foreign language from 25 different language backgrounds. Learner corpora can be used to study language acquisition, and to develop pedagogical tools and strategies for teaching English as a second or foreign language.

Many new corpora have been created in the area of language change. One of the earlier historical corpora, The Helsinki Corpus, contains 1.5 million words representing Old English to Early Modern English. Parsed versions of part of this corpus are now available and included in the Penn-Helsinki Parsed Corpus of Middle English and the Penn-Helsinki Parsed Corpus of Early Modern English. Historical corpora of English span many periods and include different types of English. The Dictionary of Old English Corpus is a three-million-word corpus containing all surviving Old English texts. The Corpus of Early English Correspondence consists of a collection of corpora containing various types of correspondence written in the fifteenth to seventeenth centuries.

There have been many changes in the ways corpora are created and analyzed. First generation corpora, such as the Brown and LOB Corpora, were relatively short (one million words in length) and contained brief samples (2,000 words) divided into different types of written English (e.g. learned writing, newspaper articles and editorials, fiction). The limited scope and length of these corpora was largely a consequence of the fact that printed texts had to be manually converted into electronic texts – a very time-consuming process. But because most texts are now available in electronic form, corpora (as noted earlier) have become considerably longer. Moreover, when initially created, the Brown Corpus, for instance, had to be loaded on to a mainframe computer for analysis, whereas many corpora such

as COCA are now available for analysis over the Web or on a home computer. Thus, the creation and dissemination of corpora have become much easier over time, resulting in a more varied selection of corpora that are considerably longer than earlier corpora.

But while printed texts can be easily included in a corpus, spoken texts still have to be manually transcribed: no voice recognition software can accurately produce a transcription because of the complexities of spoken language, particularly spontaneous conversation with its numerous hesitations, incomplete sentences, and reformulations. This is why corpora such as the Santa Barbara Corpus of Spoken American English, which is approximately 249,000 words in length, required a team of transcribers to create the corpus. Some corpora do contain transcripts of television and radio broadcasts (e.g. talk shows), but because the transcripts were created by particular broadcast agencies, their accuracy is open to question. In fact, research has shown that there can be great variability between an "in-house" transcription and what was actually said (Bednarek 2014: 54).

Considerable progress has also been made in the annotation of corpora. For instance, word-class tagging has become much more accurate. The TAGGIT Program (Greene and Rubin 1971) required 23 percent of tags to be manually disambiguated; that is, an analyst had to manually inspect a word with multiple tags and decide whether it was, for instance, a noun rather than a verb. In contrast, Manning (2011) describes ways that the Stanford Part-of-Speech Tagger, and other tagging programs as well, can be modified to achieve accuracy rates as high as 97 percent. Parsing a corpus is a more complicated undertaking because instead of analyzing individual words, a parser has to recognize larger structures, such as phrases and clauses. The British Component of the International Corpus of English (ICE-GB), a corpus containing one million words of spoken and written British English, was parsed with the Tosca/ICE parser. To ensure accuracy, the entire output of the parser was manually checked. Work with the Penn Historical Corpora has resulted in the increased accuracy of the syntactic parsing of corpora.

Other corpus annotation is used to mark additional features of texts. In a corpus of spoken dialogues, for instance, it is necessary to include tags that identify who is speaking or which sections of speaker turns contain overlapping speech. Some corpora, such as the Santa Barbara Corpus of Spoken American English, are prosodically transcribed and contain detailed features of intonation, such as pitch

contours, pauses, and intonation boundaries (cf. www.linguistics.ucsb.edu/research/transcription for further information on the transcription symbols used). Other corpora have been semantically annotated. The FrameNet Project has created corpora containing various types of semantic tags, which mark features of what are called "semantic frames" (https://framenet.icsi.berkeley.edu/fndrupal/about). For instance, the *Committing Crime* frame contains a "perpetrator" (such as a suspect) and a "crime" (such as a felony).

The main advantage of annotation is that it can greatly enhance the kinds of analyses that can be conducted on a corpus. In a tagged corpus, various kinds of lexical categories, such as proper nouns or modal auxiliaries, can be easily retrieved. In a purely lexical corpus (i.e. a corpus containing just the text), only individual words (or sequences of words) can be searched for. But even in a lexical corpus, the numerous concordancing programs that are currently available can greatly facilitate searches. The program AntConc permits searches of, for instance, the verb *walk* with all of its verb declensions (*walks*, *walking*, *walked*) (www.laurenceanthony.net/software/antconc//). However, the program ICECUP, which comes with ICE-GB, is much more powerful because it can search for grammatical categories (e.g. noun phrases) and is not restricted to lexical items.

This brief overview of the current state of corpus-based research is explored in detail in this second edition of *English Corpus Linguistics*. The book is divided into four primary chapters.

Chapter 1: The Empirical Study of Language

This chapter focuses on the empirical basis of corpus linguistics. It describes how linguistic corpora have played an important role in providing corpus linguists with linguistic evidence to support the claims they make in the particular analyses of language that they conduct. The chapter opens with a discussion of how to define a corpus, and then traces the history of corpus linguistics, noting, for instance, that corpus-based research began as early as the fifteenth century, when biblical concordances were created based on passages from the Bible. Current conceptions of corpus linguistics started with the creation of the Quirk Corpus (which contained print samples of spoken and written English) in 1955 at the Survey of English Usage at University College London. This was followed (in the 1960s) by

the Brown Corpus (which contains 2,000 word samples of various types of edited written American English).

Major research centers continued the development. At Lancaster University, one of the first major part-of-speech tagging programs, the CLAWS Tagger, automated the insertion of part-of-speech tags (e.g. noun, verb, preposition) into computer corpora. At Birmingham University, John Sinclair oversaw not just the creation of corpora but the development of their use to serve as the basis of dictionaries. But as corpus-based research began to expand during this period, its emergence during the dominance of generative grammar in the 1960s created quite a controversy, since to some linguists, corpus linguistics reeked of behaviorism. But as much research in corpus linguistics has demonstrated, empirical data can both enhance and support the claims made in linguistic analyses.

The chapter concludes with a description of the many different areas of linguistics (e.g. lexicography and sociolinguistics) that have benefited from the use of linguistic corpora, followed by a linguistic analysis illustrating that corpus-based methodology as well as the theory of construction grammar can provide evidence that appositives in English are a type of construction.

Chapter 2: Planning the Construction of a Corpus

This chapter describes both the process of creating a corpus as well as the methodological considerations that guide this process. It opens with a discussion of the planning that went into the building of four different types of corpora: the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), the Corpus of Early English Correspondence (CEEC), and the International Corpus of Learner English (ICLE). The structure of each of these corpora is discussed: their length, the genres that they contain (e.g. academic writing, fiction, press reportage, blogs, spontaneous conversations, scripted speech) as well as other pertinent information.

Subsequent sections discuss other topics relevant to planning the building of a corpus, such as defining exactly what a corpus is. Should, for instance, corpora containing samples taken from the Web be considered legitimate corpora, especially since the content of such corpora is sometimes unknown? Although this is an open question, one section of the chapter contains an analysis of web data

that precisely specifies the most common registers found in the webpages analyzed.

Other sections of the chapter focus on additional issues relevant to planning the construction of a corpus, such as how to determine the appropriate size of a corpus and the length of particular texts that the corpus will contain (complete texts versus shorter samples from each text, e.g. 2,000 words); how to select the particular genres to be included in a corpus (e.g. press reportage, technical writing, spontaneous conversations, scripted speech); and how to ensure that the writers or speakers analyzed are balanced for such issues as gender, ethnicity, and age.

Chapter 3: Building and Annotating a Corpus

This chapter focuses on the process of creating and annotating a corpus. This involves not only collecting data (speech and writing) but encoding it: transcribing recorded speech, for instance, as well as adding annotation to it, such as markup indicating in a conversation when one person's speech overlaps another speaker's, and in writing where such features as paragraph boundaries occur in written texts.

While written texts are relatively easy to collect – most writing is readily available in digital formats – speech, especially spontaneous conversations, has to be transcribed, though, as will be discussed in the chapter, voice recognition software has made some progress in automating the process for certain kinds of speech, such as monologues. Other stages of building a corpus are also discussed, ranging from the administrative (how to keep records of texts that have been collected) to the practical, such as the various ways to transcribe recordings of speech.

The chapter concludes with a detailed description of various kinds of textual markup and linguistic annotation that can be inserted into a text. Topics discussed include how to create a "header" for a particular text. Headers contain various kinds of information about the text. For instance, for written texts, the header would include such information as the title of the text; the author(s); if published, where it was published; and so forth. Other textual markup is internal to the text, and in a spoken text would include such information as speaker IDs, and the beginnings and ends of overlapping speech.

Chapter 4: Analyzing a Corpus

This chapter describes the process of analyzing a completed corpus, with an emphasis on quantitative and qualitative research methodologies along with sample corpus analyses that illustrate the direct application of these methodologies in corpus-based research. The chapter also contains discussions of corpus tools, such as concordancing programs, that can facilitate the analysis of corpora by retrieving relevant examples for a given corpus analysis, indicating their overall frequency, and specifying (depending upon the program) the particular genres in which they occur.

The first section of the chapter contains a discussion of what is termed "Trump Speak," the unique brand of language that Donald Trump uses. In addition to illustrating how a primarily qualitative corpus analysis is conducted, the analysis of Trump's speech provides a basis for describing the various steps involved in conducting a corpus analysis, such as how to frame a research question, find suitable corpora from which empirical evidence can be obtained to address the research question, and so forth. The analysis of Trump Speak also contains a discussion of how to use concordancing programs to obtain, for instance, information on the frequency of specific constructions in a corpus as well as relevant examples that can be used in subsequent research conducted on a particular topic.

The remainder of the chapter focuses on more quantitatively based corpus research, such as Douglas Biber's work on multi-dimensional analysis and the specific statistical analyses he used to determine the register distributions of a range of different linguistic constructions, for instance, the higher frequency of pronouns such as *I* and *you* in spontaneous conversations than in scientific writing. Descriptive statistics (e.g. the chi-square statistic) are illustrated with a detailed quantitative analysis of the use of a stigmatized linguistic construction, the pseudo-title (e.g. *former president Bill Clinton*), which is derived from an equivalent or appositive (*a former president, Bill Clinton*) and found mainly in newspapers. Its usage is analyzed in various regional newspapers included in various components of ICE, such as the United States, Great Britain, and Singapore.

At the end of each of these chapters are exercises that allow for the practical application of the various topics covered in the chapters. In addition, there is an Appendix that contains a listing of all corpora

discussed in the text with brief descriptions of their content as well as links to where further information on the corpora can be obtained.

In short, the goal of the second edition of *English Corpus Linguistics* is to provide a comprehensive description of all aspects of corpus linguistics, ranging from how to build a corpus to how to analyze a finished corpus. In addition, sample corpus analyses are included to illustrate the many theoretical and practical applications that linguistic corpora have.

Acknowledgments

As any writer knows, there are many people behind the scenes without whose help a book would never have been written. Such is the case with this book.

I am particularly grateful to Merja Kytö, who read an entire draft and whose insightful and judicious comments greatly improved this book.

I would also like to thank Helen Barton and Isabel Collins of Cambridge University Press for promptly replying to my many questions and for expertly taking me through the production process for the book. Their help was immeasurable.

Rachel La Russo and Minh Nghia Nguyen, doctoral students in the Applied Linguistics Department at the University of Massachusetts, Boston, were assiduous in working on the bibliography. I could not have done without their help. Robert Sigley provided me with very helpful comments on the section of chapter 4 dealing with the statistical analysis of pseudo-titles.

Finally, I wish to thank my wife, Elizbeth Fay. She kept me on track with the book through the years, gently reminding me that I needed to work on my book rather than watch yet another sporting event. And although her area of specialty is not linguistics, she helped me immensely with the many questions I had about particular topics for different chapters. Most important was her constant support.

1 The Empirical Study of Language

In an interview published in 2000, Noam Chomsky was asked "What is your view of Modern Corpus Linguistics?" His reply was, "It doesn't exist" (Aarts 2001: 5). In a talk he gave in 2011, he provides a further negative assessment of corpus linguistics, commenting:

> pretty soon you'll be able to feed the data into the computer and everything will come out. In fact, there's a field now called corpus linguistics which essentially is the same thing except that they put in the word "Bayesian" every few sentences [laughter] (Parallel Domains Workshop in honor of Jean-Roger Vergnaud, Department of Linguistics, University of Southern California, www.youtube.com/watch?v=PbK71FdF2qc)

Most individuals doing corpus-based research could probably not care less what Chomsky thinks about the research they conduct: Corpus linguistics is a firmly established area of linguistic inquiry with research being conducted in many different academic disciplines. But Chomsky's comments do provide a useful context for exploring how linguistic analyses were conducted prior to the Chomskyan revolution in the 1950s and following it. Specifically, definitions of what constitutes "empirical" evidence for linguistic analysis have shifted back and forth during this period to the point that there are now many linguists from the era of generative grammar who regularly use corpora in their current research.

To explore the role that corpora play as sources of linguistic evidence, this chapter opens with a discussion of how a linguistic corpus is actually defined, and then continues with a historical overview of the development of corpus-based research, beginning with a discussion of pre-electronic corpora; that is, corpora that were manually collected and compiled and that served as the basis of concordances and reference grammars. While pre-electronic corpora were generally based entirely on printed written texts, this changed in 1959 when Randolph Quirk at University College London

established a corpus containing spoken as well as written English, called the "Quirk" or "Survey of English Usage (SEU) Corpus." This corpus established a methodology for corpus creation and analysis that has continued until the present. Ironically, the corpus was created around the time when generative grammar completely changed the shape of linguistics. In particular, introspection replaced the collection of data as the basis of linguistics analyses, pitting the "armchair linguist," as Fillmore (1992) characterizes the introspectionist, against the corpus linguist, for whom "real" data is imperative.

In 1964, the Brown Corpus, one of the earliest computerized corpora, ushered in the "electronic" era of corpus linguistics. This corpus contained one million words of edited written American English divided into 500 samples representing different types of writing (such as fiction, technical prose, and newspaper reportage). Each sample was 2,000 words in length, enabling valid comparisons between the different registers in the corpus. The Brown Corpus was extremely important because it provided a catalyst for the many computer corpora that will be discussed throughout this book.

But as corpus linguistics developed alongside the dominant paradigm of generative grammar, obvious differences and disputes resulted. Fillmore (1992: 35) characterizes the hostility as one of different goals and values:

> The corpus linguist says to the armchair linguist, "Why should I think what you tell me is true?", and the armchair linguist says to the corpus linguist, "Why should I think what you tell me is interesting?"

However, as the example corpus analysis in the final section of the chapter demonstrates, conducting "interesting" linguistic research based on "real" samples of language are not necessarily mutually exclusive activities: a corpus can serve as a test bed for theoretical claims about language; in turn, theoretical claims can help explain patterns and structures found in a corpus.

1.1 Defining a Corpus

Although arriving at a definition of a linguistic corpus may seem like a fairly straightforward process, it is actually a more complicated undertaking than it initially appears. For instance, consider the Microsoft Paraphrase Corpus. This corpus contains 5,800 sentence pairs. One sentence of each pair was obtained from various

websites covering the news. The second sentence contains a paraphrase of the first sentence by a human annotator. For each sentence being paraphrased, information is given about its author and the particular news source from which the sentence was obtained. While it is quite common for corpora to contain meta-information about the data that they contain, should a collection of unrelated sentences and associated paraphrases be considered a corpus?

One way to answer this question is to examine the guidelines established by the Text Encoding Initiative (TEI) for the encoding of electronic texts, including linguistic corpora (www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html). Because corpora are fairly diverse in size, structure, and composition, the TEI provides a general definition of a corpus.

Several points mentioned in the guidelines are directly relevant to whether or not the Microsoft Paraphrase Corpus (MPC) fits the definition of a corpus. First, a corpus needs to be compiled "according to some conscious set of design criteria." The MPC fits this criterion, as the kinds of language samples to be included in the corpus were carefully planned prior to the collection of data. Second, a corpus can contain either complete texts (e.g. a collection of newspaper articles) or parts of texts (e.g. 500-word samples from various newspaper articles). The MPC satisfies this criterion too: each sentence in the corpus was selected from some larger text of which it is a part. The paraphrase of each sentence makes the MPC similar to a parallel corpus – a type of corpus that typically contains sentences from a whole text, with each individual sentence translated into a different language. The major difference is that parallel corpora typically contain samples from larger texts (rather than single sentences), with each sentence in the text translated into a particular language.

A third TEI guideline for defining a corpus is more problematic for the MPC. According to this guideline, a corpus is a type of text that is regarded "as *composite*, that is, consisting of several components which are in some important sense independent of each other" (www.tei-c.org/release/doc/tei-p5-doc/en/html/DS.html). It is certainly the case that the individual sentences in the MPC are individual entities. The sentences are, as the guidelines continue, "a subdivision of some larger object." But can each sentence in the MPC "be considered as a text in its own right"? (www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html).

But whether this question is answered yes or no, it is perhaps more useful, as Gilquin and Gries (2009: 6) comment, to view differing

kinds of corpora as occurring on a scale from most to least prototypical. For them, a prototypical corpus would be:

(1). machine-readable
(2). representative
(3). balanced
(4). the result of communication occurring in a "natural communicative setting."

A "machine-readable corpus" is a corpus that has been encoded in a digital format. For a corpus to be "representative," it must provide an adequate representation of whatever linguistic data is included in the corpus. For instance, a general corpus of newspaper editorials would have to cover all the different types of editorials that exist, such as op-ed pieces and editorials produced by an editorial board. For a general corpus of editorials to be "balanced," the editorials would have to be taken from the various sources in which the editorials are found, such as newspapers, magazines, perhaps even blogs. Finally, the speech or writing included in a corpus has to have been produced in a "natural communicative setting." A corpus of spontaneous conversations would need to contain unrehearsed conversations between two or more individuals in the types of settings in which such conversations are held, such as the family dinner table, a party, a chat between two friends, and so forth.

The more a corpus satisfies these four criteria, the more prototypical it would be. Gries and Gilquin (2009: 6) list the Brown Corpus (discussed in detail later in this chapter) as highly prototypical. This corpus was designed to represent edited written American English and does so by containing one million words divided into 2,000-word samples representing a broad range of types of writing, such as press reportage, editorials, and reviews; fiction; skills and hobbies; religion; and so forth. In contrast, the Penn Treebank corpus is less prototypical. Instead of containing a range of different genres of English, it consists of a heterogeneous collection of texts (totaling 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.

Obviously, the MPC would be a less prototypical corpus as well. While it is available in a machine-readable format, it is too short to be

representative of the types of sentences occurring in news stories, and whether the random sampling of sentences produced a balanced corpus is questionable. Because the sentences included in the corpus were taken from news stories, the sentences did occur in a natural communicative setting. Arguably, the translations were created in a natural communicative setting too, as a translator is often called upon to translate bits of language.

Corpora vary in size and composition largely because they have been 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 (1991b) 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 (1991b) compared the relative frequency of these modals in four genres of Australian English: press reportage, conversation, learned prose, and parliamentary debates. Collins (1991b) 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 (1991b) 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 more fact-based 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 (1991b) 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.

1.2 Pre-electronic Corpora

While modern-day corpora are closely linked to current technology (e.g. machine-readable corpora and software such as concordancing programs), it is important to note that the linguistic analysis of texts has a long history. Even though what are termed "pre-electronic corpora" were obviously not machine-readable, they were nevertheless quite influential on the development of corpus linguistics. These corpora were used for two primary purposes: to create concordances based mainly on the Bible (in many languages in addition to English), and to provide examples as the basis of prescriptive and scholarly grammars of English.

Kennedy (1998: 13) comments that biblical concordances represent "the first significant pieces of corpus-based research with linguistic associations." These concordances were written in Latin, Greek, Hebrew, and English and included Cardinal Hugo's Concordance, a Latin concordance of the Bible written in the thirteenth century; a Hebrew Concordance written by Rabbi Mordecai Nathan in the fifteenth century; and two English Concordances: John Marbeck's in the fifteenth century, and Alexander Cruden's in the eighteenth century (Keay 2005: 33–4).

Of these concordances, Cruden's stands out as the most ambitious and comprehensive. At approximately 2,370,000 words in length, it is longer than the Bible itself (Keay 2005: 29) and took a surprisingly short period of time to write. As Fraser (1996) notes, while "it had taken the assistance of 500 monks for [Cardinal] Hugo to complete his concordance of the Vulgate," Cruden took only two years to complete his concordance, working 18 hours on it every day. As Cruden describes in the introduction to the first edition in 1737, the concordance consists of three parts. Parts I and II contained an index of common and proper nouns, respectively, in the Old and New Testaments; Part III contained an index of words from books in the Bible that are, in Cruden's words, "Apocryphal"; that is, not universally accepted as legitimate parts of the Bible.

His ANNOINTED.
1 *Sam.* 2:10 exalt horn of *his a.*
12:3 against the L. and *his a.*
5 the L. and *his a.* is witness
2 *Sam.* 22:51 showeth mercy to
his a. Ps. 18:50
Ps. 2:2 and against *his a.*
20:6 the Lord saveth *his a.*
28:8 saving strength of *his a.*
Is. 45:1 saith L. to *his a.* to C.

Figure 1.1 Concordance entry for *his anointed*

Cruden's Concordance is lengthier than the Bible because he included entries not just for individual words but for certain collocations as well. In addition, Cruden does not lemmatize any of the entries, instead including separate entries for each form of a word. For instance, he has separate entries for *anoint*, *anointed*, and *anointing* as well as *his anointed*, *Lord's anointed*, and *mine anointed*. For each entry, he lists where in the Bible the entry can be found along with several words preceding and following the entry. Figure 1.1 contains a sample entry for *his anointed*.

To create the concordance, Cruden had to manually alphabetize each entry by pen on various slips of paper – an enormous amount of work. As Keay (2005: 32) notes, the letter *C* had "1019 headerwords ... of which 153 start with 'Ca'." The concordance was assembled not out of Cruden's interest in language but as a way of helping people gain easy access to the Bible.

The development of biblical concordances was followed in subsequent years by the creation of concordances of more literary texts. For instance, Heenan (2002: 9) describes the creation of a concordance of Chaucer's works, a project that began in 1871 and that consisted of a team of volunteers who were assigned sections of texts and required "to note variant spellings for each word, the definition of each word, its inflectional form, and the rhyming relationships for the final word in every line." Because the project required manual analysis, it did not appear in print until 1927 (see Tatlock 1927, Tatlock and Kennedy 1963).

Corpora served as the basis of many of the descriptively oriented grammars of English written in the late nineteenth and early to mid-twentieth centuries by individuals such as George Curme, Otto

Jespersen, Hendrik Poutsma, Henry Sweet, and Charles Fries. Not all grammarians of this period based their discussions on examples taken from a corpus. For instance, Henry Sweet's (1891–98) *A New English Grammar* is based entirely on invented examples to illustrate the grammatical categories under discussion. However, one of the more famous grammars of this era, Otto Jespersen's (1909–49) seven-volume *A Modern English Grammar on Historical Principles*, is based exclusively on examples taken from an extensive collection of written English that Jespersen consulted for examples. In fact, Jespersen may well have been the first linguist to comment directly on the merits of describing grammar on the basis of real rather than made-up examples:

> With regard to my quotations, which I have collected during many years of both systematic and desultory reading, I think that they will be found in many ways more satisfactory than even the best made-up examples, for instance those in Sweet's chapters on syntax. Whenever it was feasible, I selected sentences that gave a striking, and at the same time natural, expression to some characteristic thought; but it is evident that at times I was obliged to quote sentences that presented no special interest apart from their grammatical peculiarities. (p. vi)

Jespersen's corpus is extensive and consists of hundreds of books, essays, and poems written by well- and lesser-known authors (Vol. VII, pp. 1–40). Some of the better-known authors include Alden, Austen, Churchill, Darwin, Fielding, Hemingway, Huxley, Kipling, Locke, Mencken, Priestley, Shelley, Walpole, Wells, and Virginia Wolfe. As this list of names indicates, the writing represented in Jespersen's corpus covers a range of different written genres: fiction, poetry, science, and politics.

Reading Jespersen's description of grammatical categories, one can see that the examples he includes both shape and illustrate the points he makes. Unlike prescriptive grammarians such as Robert Lowth, Jespersen does not bring to his discussion rigidly held preconceptions of how English should be spoken and written. Instead, he uses the data in his corpus as a means of describing what the language is really like. In this sense, Jespersen is an important early influence on how descriptions of English grammar should be conducted.

A typical entry will be preceded by general commentary by Jespersen, with perhaps a few invented sentences included for purposes of illustration, followed by often lengthy lists of examples from his corpus to provide a fuller illustration of the grammatical point

being discussed. For instance, in a discussion of using a plural third person pronoun such as *they* or *their* to refer back to a singular indefinite pronoun such as *anybody* or *none*, Jespersen (Vol. II, p. 137) notes that numerous disagreements of this nature have arisen because of "the lack of a common-number (and common-sex) form in the third-personal pronoun." He then includes a quote from an earlier work of his, *Progress in Language* (published in 1894), in which he argues that using the generic *he* in a tag question such as *Nobody prevents you, does he?* "is too definite, and *does he or she?* too clumsy." He adds that using a plural pronoun in such a construction is in some cases "not wholly illogical; for *everybody* is much the same thing as *all men*," though he notes that, "this explanation will not hold good" (p. 138) for all instances of such usages. He then very exhaustively illustrates just how widespread this usage exists, giving extensive lists of examples for each of the indefinite pronouns, a sampling of which is as follows:

God send *euery one their* harts desire (Shakespeare, *Much Ado About Nothing* III 4.60, 1623)

Each had *their* favourite (Jane Austen, *Mansfield Park*, 1814)

If *anyone* desires to know . . . *they* need only impartially reflect (Percy Bysshe Shelley, *Essays and Letters*, 1912)

Now, *nobody* does anything well that *they* cannot help doing (John Ruskin, *The Crown of Wild Olive*, 1866)

Jespersen even documents instances of plural pronouns with singular noun phrases as antecedents, noting that these noun phrases often have "generic meaning" (Vol. II, p. 495):

Unless *a person* takes a deal of exercise, *they* may soon eat more than does them good (Herbert Spencer, *Autobiography*, 1904)

As for *a doctor*—that would be sinful waste, and besides, what use were *they* except to tell you what you knew? (John Galsworthy, *Caravan*, 1925)

Of course, Jespersen is not the first English grammarian to document uses of *they* with singular antecedents. Curzan (2003: 70–3) notes that such usages can be found as far back as Old English, particularly if the antecedent is a noun phrase consisting of nouns conjoined by *or* (e.g. Modern English *If a man or a woman want to get married, they must get a marriage license*). Most contemporaries of Jespersen's, she continues (pp. 73–9), treated the construction from a prescriptive point of view, in many cases insisting that generic *he* be preferred

over *they*, or that *they* be relegated to colloquial usage. And while Curzan (2003: 76) correctly observes that there is "a hint of prescriptivism" in Jespersen's discussion when he comments that using a plural pronoun with a singular antecedent "will not hold good" in all instances, Jespersen's treatment of such constructions nevertheless foreshadows the methodology now common in most corpus analyses: what occurs in one's corpus shapes the grammatical description that results.

1.2.1 The Transition from Pre-electronic to Electronic Corpora

The most ambitious pre-electronic corpus, the Quirk Corpus, served as a model for the modern-day electronic corpus. It was created at the Survey of English Usage (SEU) as part of other research being done there on the study of the English language. Although it now exists in digital form, the Quirk Corpus was originally an entirely "print" corpus. Its creation began in 1955, with final and completed digitization occurring in 1985 (www.ucl.ac.uk/english-usage/about/history.htm). The corpus totals one million words in length and contains 200 samples of spoken and written English, with each sample totaling 5,000 words in length (Greenbaum and Svartvik 1990: 11).

The idea behind the corpus was to provide as broad a representation as possible of the different types of spoken and written English that exist (Greenbaum and Svartvik 1990: 13). For instance, the spoken part of the corpus contained samples of dialogues and monologues. Within dialogues, there were conversations and public discussions. Conversations were either face-to-face or over the telephone, and were either surreptitiously or non-surreptitiously recorded – a practice that for legal and ethical reasons no longer exists. The written part of the corpus contained a likewise diverse sampling of written texts representing various printed and non-printed texts prepared to be spoken (e.g. speeches and news broadcasts).

Because the corpus was completely orthographic, it had to be prepared in a manner making it accessible to researchers. The spoken texts were transcribed and annotated with "a sophisticated marking of prosodic and paralinguistic features," and the entire corpus was "analysed grammatically" (Greenbaum and Svartvik 1990: 12). Researchers wishing to use the corpus had to travel to the Survey of English usage and conduct their grammatical analyses based on citation slips stored in file drawers. For instance, Meyer's (1987)

study of apposition is based on an analysis of all citation slips marked as "appositives," constructions such as *the author, a noted Romanticist*, and *the president of the company*, *Joyce Freeman*. From these slips, Meyer was able to find examples, classify the different the types of appositives that occurred in the corpus, and identify the different registers in which they predominated.

While the Quirk Corpus was an entirely printed corpus, work on the computerization of corpora began during the same time period. Beginning in the 1960s at the University of Edinburgh and later at Birmingham University, John Sinclair created what would become the first computerized corpus of English: a 135,000-word corpus of conversational British English (Tognini Bonelli 2001: 52). This corpus contained transcribed samples of recorded conversations. The size of the corpus was limited by the state of technology at the time:

> 135,000 words was almost the maximum that could be comfortably stored and processed, using the programs developed at the beginning of the project, given this particular machine's capacity and the time available.
>
> (Sinclair, Jones, and Daley 2004: 20)

This corpus served as an early source of information for the work that Sinclair did on English collocations over his career.

The spoken part of the Quirk Corpus was later computerized at Lund University under the direction of Jan Svartvik. The London-Lund Corpus (LLC), as it is now known, contains a total of 100 spoken texts: 87 from the original Quirk Corpus, plus an additional 13 texts added after the project was moved in 1975 to the Survey of Spoken English at Lund University in Sweden (Greenbaum and Svartvik 1990: 14). A reduced system of transcription was used to facilitate digitization. Released in 1990, this was the first computerized corpus of spoken English to be made publicly available for use by interested researchers.

But of all the early electronic corpora, the first computerized corpus of written English, the Brown Corpus (described earlier), was really the corpus that ushered in the modern-day era of corpus linguistics. Compared with present-day corpora, this corpus is relatively small (one million words). However, for the period when it was created (early 1960s), it was a huge undertaking because of the challenges of computerizing a corpus in an era when mainframe computers were the only computational devices available. Computers during this period were large machines. As Kučera (2002: 307) notes, they had

only 50 KB of RAM. This is less internal memory than the average smart phone has. Moreover, the processing of the corpus involved keying in data on punch cards or paper tape, an endeavor that took more than a year to complete. Work on the COBUILD Project at Birmingham University was likewise constrained by a "cumbersome punched-card systems for data-storage, a method which, in its most basic form, could be dated back to the eighteenth century!" (McCarthy and O'Keeffe 2010: 5).

Analyzing early versions of the Brown Corpus was an equally complicated process because these versions were released on magnetic tape. Using the corpus was a two-step process. The tape had to first be read into a mainframe computer and then punch cards prepared to conduct searches. For instance, in his analysis of punctuation usage in the Brown Corpus, Meyer (1987) had to first prepare punch cards that enabled searches for the strings such as; *and* or*; but* to find all instances of these conjunctions occurring after a semicolon. The results were displayed on printouts, and relevant examples had to be retyped for use as examples. Nowadays, a concordancing program running on a desktop or laptop computer can retrieve such information in seconds, and relevant examples can be easily copied and pasted.

Other work on the Brown Corpus introduced additional innovations. In 1967 *Computational Analysis of Present Day American English* (Kučera and Francis 1967) was published, a book that contained information on word frequencies in the Brown Corpus. This book provided word frequency lists for the entire corpus as well as the individual genres in it. Although word frequency lists had been created previously, this was the first such list that was based on a carefully collected group of texts and that contained frequency information not just for the entire corpus but for the individual genres of the corpus. Creating this list on such a large corpus required "14 hours of continuous dedicated processing on a million-dollar computer with the aid of six tape drives" (Kučera 2002: 307).

The Brown Corpus was also the first corpus to be lexically tagged; that is, each word was assigned a part of speech designation (e.g. the tag DO for the verb *do* or DOD for the past tense form *did*). All 77 tags were assigned to each word in the corpus by a computer program designed at Brown University called "TAGGIT" (Greene and Rubin 1971). This program was not as accurate as current taggers, with 23 percent of the tags ultimately requiring manual disambiguation (Francis 1979). Nevertheless, as Hockey (2000:

94–5) comments, the tagged version of the Brown Corpus was an extremely important achievement "not only because it was the first [tagged corpus] and laid down the methodological groundwork for word class tagging but because it provided an accurate tagged corpus which could be used as a reference by other tagging systems." The tagged corpus served as the basis of a second book, *Frequency Analysis of English Usage: Lexicon and Grammar* (Francis and Kučera 1982), which listed word frequencies for grammatical categories (e.g. *talk* used as a noun as well as a verb).

The early influences on corpus linguistics discussed in this section do not exhaust the many other factors that affected the current state of corpus linguistics. One of the key developments, as McCarthy and O'Keeffe (2010: 5) note, "was the revolution in hardware and software in the 1980s and 1990s that really allowed corpus linguistics as we know it to emerge." Advances in technology have resulted in, for instance, web-based corpora and textual analysis software, such as concordancing programs, that are fast and can be run on desktop and laptop computers. In their overview of the history of corpus linguistics, McEnery and Hardie (2012) note among many influences important research centers where work on corpus linguistics was conducted. For instance, at the UCREL Center at Lancaster University, the CLAWS Tagger was developed under the direction of Geoffrey Leech and Roger Garside; it is regarded as "the first viable automated part-of-speech tagging program" (McEnery and Hardie 2012: 77). At Birmingham University, John Sinclair headed the COBUILD research unit. This unit not only developed many corpora, such as the Bank of English Corpus, but conducted research on lexicology and collocations culminating in the publication of numerous dictionaries based on corpora as well as important theoretical advances in lexico-grammar (McEnery and Hardie 2012: 79–80).

1.3 Corpus Linguistics in the Era of Generative Grammar

At the time when the Brown Corpus was 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 linguistics 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.

This attitude was largely a consequence of the conflict between what Chomskyan and corpus linguists considered "sufficient" evidence for linguistic analysis. In short, corpus linguists were grouped into the same category as the structuralists – "behaviorists" – that Chomsky had criticized in the early 1950s as he developed his theory of generative grammar. For Chomsky, the mere "description" of linguistic data was a meaningless enterprise. It was more important, he argued, that grammatical descriptions and linguistic theories be evaluated in terms of three levels of "adequacy": *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," that is, the omission of a subject pronoun 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 is 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.

Chomsky also distinguishes 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 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 to 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."

Corpus linguists also challenge the clear distinction made in generative grammar between competence and performance, and the notion that corpus analyses are overly descriptive rather than theoretical. Leech (1992: 108) argues that both these criticisms are overstated: the distinction between competence and performance is not as great as is often claimed, "since the latter is the product of the former." Consequently, what one discovers in a corpus can be used as the basis for whatever theoretical issue one is exploring. In addition, all of the criteria applied to scientific endeavors can be satisfied in a corpus study, since corpora are excellent sources for verifying the falsifiability, completeness, simplicity, strength, and objectivity of any linguistic hypothesis (Leech 1992: 112–13).

Corpus-based studies also expand the range of data considered for analysis beyond the linguist's intuitions, and ensure the accuracy of any generalizations that are made. For instance, in a discussion of Pollard and Sag's (1994) treatment of verb subcategorization in English, Manning (2002: 299) notes mistakes in the analysis because of an over-reliance on data gathered introspectively. In their analysis, Pollard and Sag (1994) claim that while the verb *consider* can take a predicative *to*-complement in (a), it cannot take a predicative *as*-complement in (b):

(a) We consider Kim to be an acceptable candidate.
(b) *We consider Kim as an acceptable candidate.

However, Manning (2002: 299) notes that he found many examples such as (c) in the *New York Times*, where *consider* clearly takes an *as*-complement:

(c) The boys consider her as family and she participates in everything we do.

Manning (2002: 299) comments that if examples such as (c) were rare and uncommon, then "Pollard and Sag got that one particular fact wrong." But he found other cases where Pollard and Sag ignored relevant data.

But while corpora are certainly essential linguistic resources, it is important to realize that no corpus, regardless of its size, will contain every relevant example of a particular linguistic construction or be able to provide a fully complete picture of how the construction is used. At best, a corpus can provide only a "snapshot" of language usage. For this reason, many have argued that corpus data should be supplemented with other means of gathering data.

Greenbaum (1984) is one of the earlier discussions of how elicitation tests can be used in conjunction with findings from a corpus analysis. Drawing upon earlier work on elicitation experiments in Greenbaum and Quirk (1970), he demonstrates how corpus material available from the spoken and written components of the Survey of English Usage Corpus can be augmented with two different types of tests: performance and judgment tests. One type of performance test, the completion test, asks subjects to perform a type of operation on a sentence. For instance, the "composition" test below asked subjects to create a sentence based on the following two words:

They/badly. . . - Complete the sentence

The purpose of this test was to determine which verbs collocate with intensifying adverbs such as *badly*. Possible answers are *need* or *want*.

In contrast, judgment tests, such as the "evaluation" test below, require subjects to assess the acceptability of sentences in various ways. In this case, subjects are asked to rank their preferences for the three sentences, from highest to lowest:

He doesn't have a car.
He hasn't a car.
He hasn't got a car.

The purpose of this test was to examine differences in negation and the use of *got* in British and American English. Thus, while speakers of American English will rank *He hasn't got a car* higher than *He hasn't a car*, speakers of British English will reverse the rankings.

Kepser and Reis (2005) frame the issue of corpus and experimental evidence more generally. They call for "a rapprochement between introspective and corpus linguists" (p. 3) and argue that linguistic evidence should be multi-faceted and involve "corpus data, introspective data, psycholinguistic data, data from computational linguistics, language acquisition data, data from the historical linguistics, and sign language data" (p. 4).

Gilquin and Gries (2009) echo Kepser and Reis' call for linguistic analyses to be based on both corpus and experimental evidence. They note that both corpus-based and experimental-based studies each have various advantages and disadvantages, and as a consequence, the two approaches can be seen as "complementary" (p. 9). For instance, one feature of a corpus that is lacking in an experiment is that a corpus contains data occurring in "natural contexts"; one advantage of an experiment is that it can be designed to study "phenomena that are too infrequent in corpora" (p. 8).

1.4 Types of Corpora

There are many kinds of corpora that have been created to fulfill the research needs of those doing corpus-based research. These corpora range from *multipurpose corpora*, which can be studied to carry out many differing types of corpus-based analyses, to *learner corpora*, which have been designed to study the types of English used by individuals (from many differing first language backgrounds) learning English as a second or foreign language. This section provides an overview of the many different types of corpora that exist. Less detail is given to corpora discussed in greater detail in subsequent chapters.

1.4.1 Multipurpose Corpora

These are corpora like the London-Lund Corpus, the British National Corpus (BNC), and the Corpus of Contemporary American English (COCA). All of these corpora contain numerous registers of speech or writing (such as fiction, press reportage, casual conversations, and spoken monologues or dialogues) representative of a particular variety of English. For instance, the London-Lund Corpus contains different kinds of spoken British English, ranging from casual conversation to course lectures. The BNC and COCA contain various samples of spoken and written British and American English, respectively.

Multipurpose corpora are useful for many different types of analyses. Because the London-Lund Corpus has been prosodically transcribed, it can be used to study various features of British English intonation patterns, such as the relationship between grammar and intonation in English. For instance, in a noun phrase such as *the leader of the country and her cabinet*, a tone unit boundary will typically occur before the conjunction *and*, marking the boundary

between two coordinated noun phrases. There is also currently in production an updated version of the London-Lund Corpus: LLC-2 (https://projekt.ht.lu.se/llc2/guide-to-the-corpus). It contains spoken texts collected between 2014–2019 that replicate the particular texts included in the original London-Lund Corpus, thus permitting comparisons between the structure of spoken British English between the original corpus and the new one.

The BNC and COCA differ in length: the BNC is 100 million words in length, whereas COCA is a monitor corpus; new texts are constantly added to it so that it currently contains one billion words. But despite the difference in length, each corpus contains many of the same registers, such as fiction, press reportage, and academic writing. While the BNC contains spontaneous conversations that have been manually transcribed, COCA is restricted to spoken registers from which published transcriptions are available. Consequently, it contains no spontaneous conversations. The BNC is a fixed corpus: its structure hasn't been changed since its creation in the 1990s, though a successor corpus, BNC2014, is now in progress (cf. http://corpora.lancs.ac.uk/bnc2014/ and also Love 2020).

The International Corpus of English (ICE) contains comparable one million-word corpora of spoken and written English representing the major national varieties of English, including English as it is spoken and written in countries such as Ireland, Great Britain, the Philippines, India, and many other varieties as well.

1.4.2 Learner Corpora

This type of corpus contains texts that represent the speech or writing of individuals who are in the process of learning a language as a second or foreign language. For instance, the International Corpus of Learner English contains samples of written English from individuals who speak English as a foreign language and whose native languages include French, German, Portuguese, Arabic, and Hungarian. Learner corpora enable researchers to study such matters as whether one's native language affects their mastery of a foreign language, in this case English. There are other learner corpora representing numerous languages other than English, including Arabic, Hungarian, and Malay, as well as projects providing various ways that learner corpora can be analyzed (for more comprehensive listing of learner corpora, see https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html).

1.4.3 Historical Corpora

The corpora discussed thus far contain examples of various types of modern English. However, there are also diachronic or historical corpora, which contain texts from earlier periods of English.

Much of the interest in studying historical corpora began with the creation of the Helsinki Corpus, a 1.5 million-word corpus of English containing texts from the Old English period (beginning in the eighth century) through the early Modern English period (the first part of the eighteenth century). Texts from these periods are further grouped into sub-periods (ranging from 70 to 100 years) to provide what Rissanen (1992: 189) terms a "chronological ladder" of development; that is, a grouping of texts from a specific period of time that can be compared with other chronological groupings of texts to study major periods of linguistic development within the English language. In addition to covering various periods of time, the texts in the Helsinki Corpus represent various dialect regions in England as well as different genres (e.g. law, homilies and sermons, fiction, and letters; for full details, see Kytö 1996). The corpus also contains sociolinguistic information on authors (e.g. age, gender) for texts written from the Middle English period onwards, since prior to this period sociolinguistic "information is too inconsistent to form a basis for any socio-historical considerations" (Rissanen 1992: 193).

To fill gaps in the Helsinki Corpus, ARCHER (A Representative Corpus of Historical English Registers) was created (cf. Biber et al. 1994; Biber and Burges 2000; see also www.projects.alc.manchester.ac.uk/archer/. Version 3.2 of ARCHER is 3.3 million words in length and contains American as well as British English. It covers the years 1650–1990, with texts divided into 50-year sub-groups. Like the Helsinki Corpus, it contains various genres of English, representing not just formal exposition (e.g. scientific writing) but personal letters and diaries and more spoken-based genres, such as sermons and fictional dialogue. Because of copyright restrictions, the number of words that can be retrieved in searches is restricted (www.projects.alc.manchester.ac.uk/archer/using-archer/).

The Penn Parsed Corpora of Historical English contains components that cover three different periods of the English language: Middle English, Early Modern English, and Modern British English:

The Penn-Helsinki Parsed Corpus of Middle English, second edition
The Penn-Helsinki Parsed Corpus of Early Modern English
The Penn Parsed Corpus of Modern British English, second edition

Each corpus has been both lexically tagged and syntactically parsed. That is, each word has been assigned a lexical tag identifying its part of speech (such as verb, noun, preposition). In addition, the larger corpora have been syntactically parsed, so that both individual words as well as structures, such as noun phrases, can be retrieved. Users of the corpus will therefore be able to study the development over time of both individual words as well as a variety of different syntactic structures.

Although FLOB (The Freiburg LOB Corpus of British English) and FROWN (The Freiburg Brown Corpus of American English) are not historical corpora in the sense that the Helsinki and ARCHER Corpora are, they do permit the study of changes in British and American English between the periods 1961–1991. Moreover, FLOB and FROWN replicate the LOB and Brown Corpora, respectively, but with texts published in the year 1991. Thus, FLOB and FROWN allow for studies of linguistic change in British and American English over a period of 30 years. Although 30 years is not a long time for language to change, studies of FLOB and FROWN have documented changes in the language during this period (cf., for instance, Mair 1995). Consequently, FLOB and FROWN are synchronic corpora on the one hand, providing resources for describing Modern English, but diachronic corpora on the other when compared with the LOB and Brown Corpora.

The two main historical corpora, the Helsinki and ARCHER Corpora, contain many different texts and text fragments covering various periods of English. There are, however, more focused historical corpora covering specific works, authors, genres, or periods. These corpora include an electronic version of *Beowulf*, "The Electronic *Beowulf*" (cf. Prescott 1997 and http://ebeowulf.uky.edu/), which is searchable online (https://ebeowulf.uky.edu/ebeo4.0/CD/main.html); the works of Chaucer and other Middle English writers (the Corpus of Middle English Prose and Verse; cf. https://quod.lib.umich.edu/c/cme/); collections of early English letters (the Corpus of Early English Correspondence; cf. Nevalainen and Raumolin-Brunberg 1996); and early Modern English tracts (the Lampeter Corpus; cf. Schmied and Claridge 1997). Kytö (2012: 1523) describes the growth of historical corpora, noting the many different types of historical corpora that currently exist, including "multipurpose corpora, specialized corpora, electronic text additions, large-scale tax collections, and electronic dictionaries."

1.4.4 The Web as Corpus

Because the Web can be searched, it can be regarded as a corpus with infinite boundaries. Initially, as Gatto (2014: 36) notes, many corpus linguists were skeptical of basing corpus analyses on data obtained from the Web:

> Why should linguists, translators, language professionals of any kind, turn their attention to a collection of texts whose content is largely unknown and whose size is difficult to measure?

However, she describes several ways in which the web can be usefully studied. For instance, it can be used as a "corpus surrogate," a source of supplemental data, or as though it were a "corpus proper" (Gatto 2014: 37).

Chapter 4 will contain a discussion of "Trump Speak": the type of language that Donald Trump used when he was president. The Web was instrumental in locating additional data for the study, since Trump is quoted extensively in newspapers and other print media, making it possible to find additional examples of his usage of language that proved very useful for the study. For instance, when Marco Rubio was running against Trump in the primaries, Trump repeatedly referred to him as "Little Marco." While this usage shows up regularly in Trump's tweets, an example from the Web containing this usage during a debate helped establish this usage as not simply restricted to tweets:

> "I will. Don't worry about it, Marco. Don't worry about it," Trump said to an applauding crowd. "Don't worry about it, little Marco, I will."

In addition, there are corpora that have been created based solely on data from the Web. For instance, the iweb Corpus is 14 billion words in length and is searchable online. It was systematically created. Using information from alexa.com, one million of the most popular websites were downloaded. These websites were then screened to ensure that they were from countries in which English is spoken as a native language: the United States, the UK, Canada, Ireland, Australia, and New Zealand. Several processes of elimination were developed to further screen the samples, resulting in a corpus based on 94,391 websites that yielded 22,388,141 webpages.

1.4.5 Parallel Corpora

A parallel corpus is a very specialized corpus. As Gatto (2014: 16) notes, "parallel corpora consist of original texts and their

translations into one or more languages." For instance, the English–Norwegian parallel corpus contains texts in both English and Norwegian that are then translated into the other language, for instance English into Norwegian and Norwegian into English (https://benjamins.com/catalog/scl.90). Such corpora are useful for translation studies, as they enable researchers to study how structures are changed after they are translated into a different language.

1.5 Uses of Corpora

The previous section provided a description of several different types of corpora. This section provides a sampling of the types of analyses that can be conducted on corpora.

1.5.1 Language Variation

Because corpus linguists are interested in studying the contexts in which language is used, modern-day corpora, from their inception, have been purposely designed to permit the study of what is termed "genre variation"; that is, how language usage varies according to the context in which it occurs. The first computer corpus, the Brown Corpus (a general-purpose corpus), contained various kinds of writing, such as press reportage, fiction, learned, and popular writing. In contrast, the MICASE Corpus is a more specialized corpus that contains various types of spoken English used in academic contexts (e.g. advising sessions, colloquia) at the University of Michigan.

Biber's (1988) study of the linguistic differences between speech and writing illustrates the potential that general-purpose corpora have for yielding significant insights into the structure of a range of different written and spoken genres of English. Using the LOB Corpus of writing and the London-Lund Corpus of speech, Biber (1988) was able to show that contrary to the claims of many, there is no strict division between speech and writing but rather that there exists a continuum between the two: certain written genres (such as fiction) contain linguistic structures typically associated with speech, whereas certain spoken genres (such as prepared speeches) contain structures more commonly associated with written texts.

Biber (1988: 13) was interested in grammatical co-occurrences such as these because, he argued, "that strong co-occurrence patterns

of linguistic features mark underlying functional dimensions"; that is, that if passives and conjuncts (e.g. *therefore* or *nevertheless*) occur together, for instance, then there is some functional motivation for this co-occurrence. The functional motivations that Biber (1988) discovered led him to posit a series of "textual dimensions." Passives and conjuncts are markers of abstract uses of language, Biber (1988: 151–4) maintains, and he places them on a dimension he terms "Abstract versus Non-Abstract Information." Low on the dimension are two types of spoken texts that contain relatively few abstractions: face-to-face conversations and telephone conversations. Biber's work on multidimensional analysis is discussed in greater detail in Chapter 4.

While genre or register variation focuses on the particular linguistic constructions associated with differing text types, sociolinguistic variation is more concerned with how various sociolinguistic variables, such as age, gender, and social class, affect the way that individuals use language. One reason that there are not more corpora for studying this kind of variation is that it is tremendously difficult to collect samples of speech, for instance, that are balanced for gender, age, and ethnicity. Moreover, once such a corpus is created, it is less straightforward to study sociolinguistic variables than it is to study genre variation. To study press reportage, for instance, it is only necessary to take from a given corpus all samples of press reportage, and to study within this sub-corpus whatever one wishes to focus on. To study variation by gender in, say, spontaneous dialogues, on the other hand, it becomes necessary to extract from a series of conversations in a corpus what is spoken by males as opposed to females – a much more complicated undertaking, since a given conversation may consist of speaker turns by males and females distributed randomly throughout a conversation, and separating out who is speaking when is neither a simple nor straightforward computational task. Additionally, the analyst might want to consider not just which utterances are spoken by males and females but whether an individual is speaking to a male or female, since research has shown that how a male or female speaks is very dependent upon the gender of the individual to whom they are speaking.

But despite the difficulties of creating a truly balanced corpus, designers of the BNC made concerted efforts to produce a corpus that was balanced for such variables as age and gender, and that was created so that information on these variables could be extracted by various kinds of software programs. Prior to the collection of

spontaneous dialogues in the BNC, calculations were made to ensure that the speech to be collected was drawn from a sample of speakers balanced by gender, age, social class, and dialect region. Included within the spoken part of the BNC is a sub-corpus known as the Corpus of London Teenage English (COLT). This part of the corpus contains a valid sampling of the English spoken by teenagers from various socioeconomic classes living in differing boroughs of London.

To enable the study of sociolinguistic variables in the spoken part of the BNC, each conversation contains a file header, a statement at the start of the sample providing such information as the age and gender of each speaker in a conversation. A software program, SARA (SGML-Aware Retrieval Application), was designed to read the headers and do various analyses of the corpus based on a pre-specified selection of sociolinguistic variables. Using SARA, Aston and Burnard (1998: 117–23) demonstrate how a query can be constructed to determine whether the adjective *lovely* is, as many have suggested, used more frequently by females than males. After using SARA to count the number of instances of *lovely* spoken by males and females, they confirmed this hypothesis to be true.

Other corpora have been designed to permit the study of sociolinguistic variables as well. In the British component of the International Corpus of English (ICE-GB), ethnographic information on speakers and writers is stored in a database, and a text analysis program designed to analyze the corpus, ICECUP (the International Corpus of English Corpus Utility Program) can draw upon information in this database to search by, for instance, age or gender. Even though ICE-GB is not balanced for gender – it contains the speech and writing of more males than females – a search of *lovely* reveals the same usage trend for this word that was found in the BNC.

Of course, programs such as SARA and ICECUP have their limitations. In calculating how frequently males and females use *lovely*, both programs can only count the number of times a male or female speaker uses this expression; neither program can produce figures that, for instance, could help determine whether females use the word more commonly when speaking with other females than males. And both programs depend heavily on how accurately and completely sociolinguistic variables have been annotated, and whether the corpora being analyzed provide a representative sample of the variables. In using SARA to gather dialectal information from the BNC, the analyst would want to spot check the ethnographic

information on individuals included in the corpus to ensure that this information accurately reflects the dialect group in which the individuals are classified. Even if this is done, however, it is important to realize that individuals will "style-shift": they may speak in a regional dialect to some individuals but a more standard form of the language with others. In studying variation by gender in ICE-GB, the analyst will want to review the results with caution, since this corpus does not contain a balanced sample of males and females. Software such as SARA or ICECUP may automate linguistic analyses, but they cannot deal with the complexity inherent in the classification of sociolinguistic variables. Therefore, it is important to view the results generated by such programs with a degree of caution.

More recent work has investigated sociolinguistic variables in a number of different corpora. Murphy (2010) analyzed a 90,000-word corpus that she created, the Corpus of Age and Gender, which contained two sub-corpora: the Female Adult Corpus and the Male Adult Corpus. Each sub-corpus contained samples from three age groups: individuals in their 20s, 40s, and 70s/80s (p. 32). All participants in her study spoke Irish English. Her goal in creating these corpora was to study the effects that both age and gender had on the use of language.

One topic she discussed to illustrate the role that age plays in language usage focused on the use of hedges in constructions containing verbs (*I think* and *I might*) and adverbs (*maybe* and *probably*). Table 1.1 is the one she created (adapted from Table 4.1, p. 62) to illustrate how each of the age groups used these two types of hedges

Murphy was able to draw a number of conclusions from her study. First, while the 20s and 40s age group used roughly the same number of hedges, the 20s group favored the adverb form, while the 40s group preferred the verb form. But the major difference between the two groups was that the 20s group favored the adverb form for

Table 1.1 *Adapted from Murphy (2010)*

Age group	*Verb form*	*Adverb form*	*Total*
20s	7,436	12,179	19,615
40s	11,178	8,349	19,527
70s/80s	5,099	3,091	8,190
Totals	23,713	23,619	47,332

hedges, while the 40s group preferred the verb form. As Murphy comments (p. 62): "They are conscious of being part of a group and are aware of the importance of maintaining and keeping friendships." In contrast, because the 70s/80s group "have known each other for many years… it would appear that hedging is less important for them." With friendships that have been solidified, she concludes that hedging becomes less necessary. However, she makes an important point concerning this interpretation, noting that "more qualitative examination is needed" (p. 62).

1.5.2 Lexicography

Studies of grammatical constructions can be reliably conducted on corpora of varying length. However, to obtain valid information on vocabulary items for the purpose of creating a dictionary, it is necessary to analyze corpora that are very large. To understand why this is the case, one need only investigate the frequency patterns of vocabulary in corpora, such as the one million-word BNC. In the BNC, the five most frequent lexical items are the function words *the*, *of*, *and*, *a*, *in* (https://ucrel.lancs.ac.uk/bncfreq/):

Word	*PoS*	*Freq*
the	Det	61,847
of	Prep	29,391
and	Conj	26,817
a	Det	21,626
in	Prep	18,214

The five least frequent lexical items are not five single words but rather hundreds of different content words that occur 10–15 times each in the corpus. These words include numerous proper nouns, such as *Bond* and *MacDonald*, as well as miscellaneous content words such as *bladder, dividends*, and *woodland*. These frequencies illustrate a simple fact about English vocabulary (or, for that matter, vocabulary patterns in any language): a relatively small number of words (particularly function words) will occur with great frequency; a relatively large number of words (content words) will occur far less frequently. Obviously, if the goal of a lexical analysis is to create a dictionary, the examination of a small corpus will not give the lexicographer complete information concerning the range of

vocabulary that exists in English and the varying meanings that these vocabulary items will have.

Because a traditional linguistic corpus, such as the LOB Corpus, "is a mere snapshot of the language at a certain point in time" (Ooi 1998: 55), some have argued that the only reliable way to study lexical items is to use what is termed a "monitor" corpus; that is, a large corpus that is not static and fixed but that is constantly being updated to reflect the fact that new words and meanings are always being added to English. This was the philosophy of the Collins COBUILD Project at Birmingham University in England, which produced a number of dictionaries based on two monitor corpora: the Birmingham Corpus and the Bank of English Corpus. The Birmingham Corpus was created in the 1980s (cf. Renouf 1987 and Sinclair 1987), and while its size was considered large at the time (20 million words), it would now be considered fairly small, particularly for the study of vocabulary items. For this reason, the Birmingham Corpus has been superseded by the Collins Corpus, which is 4.5 billion words in length.

To understand why dictionaries are increasingly being based on corpora, it is instructive to review precisely how corpora, and the software designed to analyze them, can not only automate the process of creating a dictionary but improve the information contained in the dictionary. A typical dictionary, as Landau (1984: 76f.) observes, provides its users with various kinds of information about words: their meaning, pronunciation, etymology, part of speech, and status (e.g. whether the word is considered "colloquial" or "non-standard"). In addition, dictionaries will contain a series of example sentences to illustrate in a meaningful context the various meanings that a given word has.

Prior to the introduction of computer corpora in lexicography, all of this information had to be collected manually. As a consequence, it took years to create a dictionary. For instance, the most comprehensive dictionary of English, the *Oxford English Dictionary* (originally entitled *New English Dictionary*), took 50 years to complete, largely because of the many stages of production that the dictionary went through. Landau (1984: 69) notes that the five million citations included in the *OED* had to be "painstakingly collected. . .subsorted. . .analyzed by assistant editors and defined, with representative citations chosen for inclusion; and checked and redefined by [James A. H.] Murray [main editor of the *OED*] or one of the other supervising editors." Of course, less ambitious

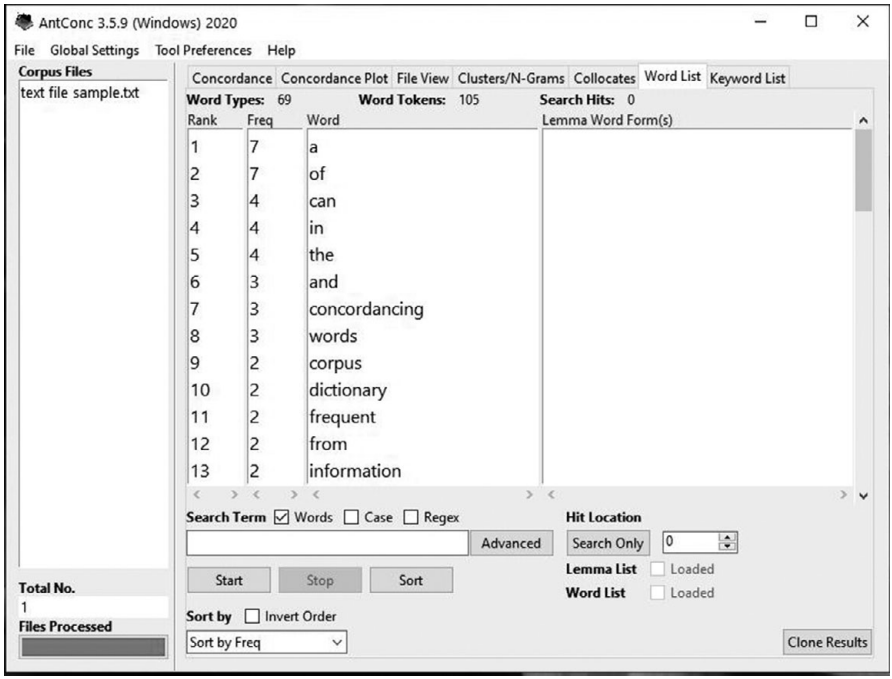

Figure 1.2 Concordance results

dictionaries than the *OED* took less time to create, but still the creation of a dictionary is a lengthy and arduous process.

Because so much text is now available in computer-readable form, many stages of dictionary creation can be automated. Using a relatively inexpensive piece of software called a concordancing program, the lexicographer can go through the stages of dictionary production described above, and instead of spending hours and weeks obtaining information on words, can obtain this information automatically from a computerized corpus. In a matter of seconds, a concordancing program can count the frequency of words in a corpus and rank them from most frequent to least frequent. Figure 1.2 is a concordancing window for some of the more frequently occurring words in this paragraph.

Note the high relative frequency of function words, the modal verb *can*, and other vocabulary items, such as *words* or *concordancing*, that are topics of the discussion in the paragraph.

In addition, some concordancing programs can detect prefixes and suffixes and irregular forms and sort words by "lemmas"; that is, words such as *runs, running*, and *ran* will not be counted as separate entries but rather as variable forms of the lemma *run*. And as

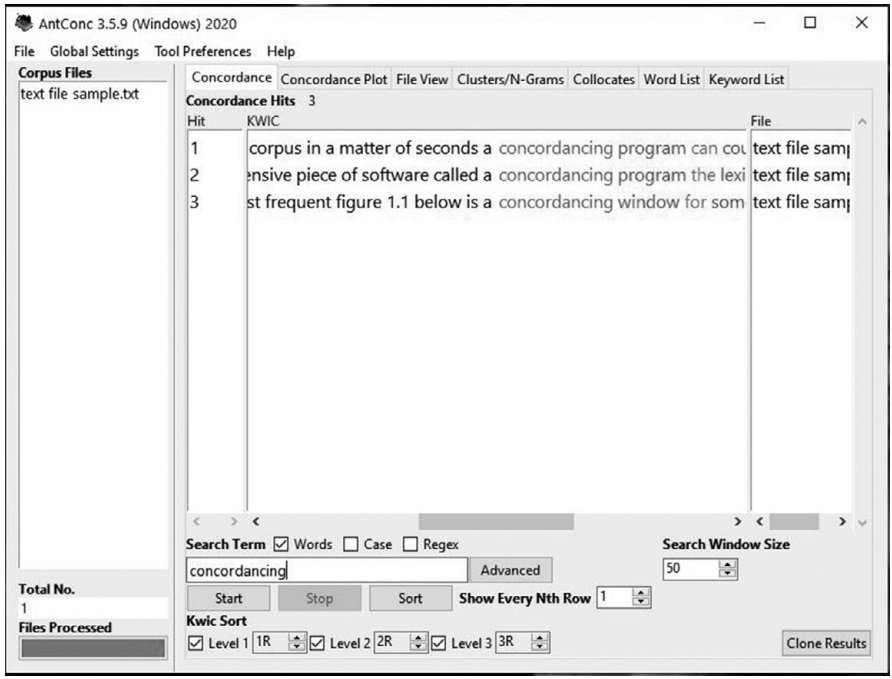

Figure 1.3 KWIC (key word in context) format

illustrated in Figure 1.3, words can be displayed in KWIC (key word in context) format, and easily viewed in the varying contexts in which the word occurs.

Furthermore, if the lexicographer desires a copy of the sentence in which a word occurs, it can be automatically extracted from the text and stored in a file, making obsolete the handwritten citation slip stored in a filing cabinet. If each word in a corpus has been tagged (i.e. assigned a tag designating its word class), the part of speech of each word can be automatically determined. In short, computer corpora and associated software have completely revolutionized the creation of dictionaries.

As Kilgarriff and Kosem (2012) note, there are currently a number of "corpus tools" available to the lexicographer that can greatly automate the retrieval of information necessary for the creation of a dictionary entry. For instance, one tool that they describe creates what are termed "word sketches." These are short descriptions of the words as well as forms and functions that collocate with a particular dictionary entry. One example they list is the verb *caress* (p. 44), and the frequent forms and functions that this particular verb collocates with.

Function	Form
Subject	*hand* (hand caresses)
Modifiers	*gently* (gently caress)
Object	*face* (caress the face)
Pronoun objects	*her/me* (caress her/me)
Conjunctions	*and/or* (caress and kiss)

For instance, *caress* collocates with the noun *hand* when *hand* is functioning as the subject of a sentence. It also collocates with the modifier *gently*. Such collocations are useful in lexicography because they specify contexts in which words occur – contexts that can help identify exactly what a particular word means.

1.6 Corpus-Based Research in Linguistics: A Case Study

Linguists from many different theoretical perspectives have discovered that corpora can be very useful resources for pursuing various research agendas. For instance, lexicographers, as noted above, have found that they can more effectively create dictionaries by studying word usage in very large linguistic corpora. Much current work in historical linguistics is now based on corpora containing texts taken from earlier periods of English – corpora that permit a systematic study of the development of English and that enable historical linguists to investigate issues that have currency in modern linguistics, such as whether males and females used language differently in earlier periods of English. Corpora have been introduced into other linguistic disciplines as well, and have succeeded in opening up new areas of research or bringing new insights to traditional research questions.

Chapter 4 ("Analyzing a Corpus") describes in detail how to conduct a corpus analysis, covering such topics as how to frame a research question for a particular corpus analysis and select the appropriate corpora for conducting the analysis. But this section demonstrates how a particular theory of language, Construction Grammar, can provide insights into the variable grammatical category of appositions in English. Such linkages between theory and practice are important because they, on the one hand, avoid the oft-made criticism of corpus linguists that they are mainly "bean counters" – linguists who fail to make connections between theory and usage – with too little

attention paid to the theoretical underpinnings of the data they are describing.

1.6.1 Appositions as Constructions

Apposition has proven to be a problematic grammatical category, largely because treatments of it disagree about which constructions should be considered appositions. For instance, most studies consider a construction such as *Geoffrey Plimpton, police commissioner* as consisting of two units in apposition. However, if the two units are reversed, a reversal possible with some but not all appositions, a very different construction, *police commissioner Geoffrey Plimpton*, results – one that some studies favoring a more expansive view of apposition consider appositional (e.g. Meyer 1992), but that those advocating a more restricted view do not (e.g. Acuña 1996). These conflicting views are each problematical. Expansive views of apposition posit a series of syntactic and semantic gradients to distinguish various "degrees" of apposition. The result is the admission of constructions such as *a person like you* into the category of apposition, a construction so different from typical examples of apposition that some argue it renders the notion of apposition almost meaningless. The latter view is so restrictive that any construction that does not contain juxtaposed noun phrases separated by an intonation boundary is not considered appositional.

In more recent work, Acuña (2006) attempts to bridge the gap between these two views. Working within the framework of construction grammar/cognitive linguistics, he views the many constructions considered appositional as occupying what he terms "appositive space," a space where the differing types of appositions are related through notions such as "family resemblance" and "prototype." Construction grammar is a particularly useful theory for describing apposition, since within construction grammar, constructions are viewed as "'vertical' structures" (Croft and Cruse 2004: 247); that is, structures that are defined not simply syntactically but semantically, phonologically, and pragmatically as well. In other words, constructions are viewed as "conventionalized pairings of form and function" (Goldberg 2006: 3): an apposition such as *the current President of the US, Barack Obama* has a particular form – two noun phrases – that are juxtaposed and co-referential and that serve a particular communicative function, in this case to identify who Barack Obama is.

To illustrate why appositions are constructions, it is helpful to examine the most frequent apposition that Meyer (1992) found in the corpora that he analyzed: the nominal apposition containing a proper noun in one of the two units (hereafter referred to as APNs), such as *Geoffrey Plimpton* occurring with *police commissioner* in the earlier cited example. These appositions also had a very restricted usage: they overwhelmingly occurred in press reportage. The form of these appositions, plus their occurrence primarily in press reportage, can be explained by examining corpus data and then linking the corpus data to the notion of "construction" in Construction Grammar.

1.6.2 The Frequency and Distributions of APNs in the Corpora

Meyer (1992: 10–34) provides a detailed description of the forms of appositions occurring in a 360,000-word corpus of spoken and written American and British English that he investigated. The spoken part of the corpus consisted of 120,000 words of spontaneous conversation (British English) taken from the London-Lund Corpus. The written part was comprised of 240,000 words of press reportage, learned prose (humanistic/scientific), and fiction divided evenly between the Brown Corpus (American English) and the Quirk Corpus (British English), which was available at the time in printed form at the Survey of English Usage, University College London.

Figure 1.4 contains the frequencies of the four principal forms of apposition that Meyer (1992: 11) found.

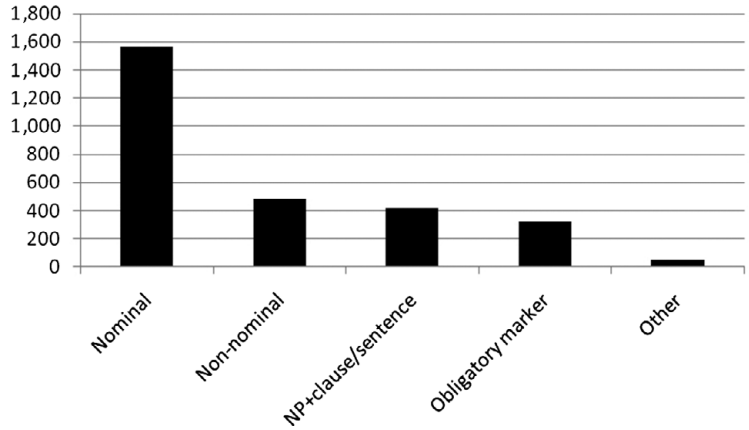

Figure 1.4 The forms of apposition (raw frequencies)

The most common type of apposition contained two juxtaposed noun phrases in apposition:

(1) *The first twenty thousand pounds, the original grant*, is committed. (London-Lund Corpus S.1.2 782–3)

Occurring far less frequently were three additional types of appositions:

(a) non-nominal appositions, such as the two adjective phrases below:

(2) when the patient closed his eyes, he had absolutely *no spatial* (that is, *third dimensional*) awareness whatsoever. (Brown J52 100–50)

One reason these two adjectives are considered in apposition is that the second unit is preceded by the marker of apposition *that is*, a marker that can precede more "canonical" appositions consisting of units that are noun phrases.

(b) a noun phrase in apposition with some kind of clausal structure, such as the *wh*-question below:

(3) The Sane Society is an ambitious work. Its scope is as broad as *the question: What does it mean to live in modern society*? (Brown J63 010–20)

Although the two units in this example are not co-referential (only noun phrases can co-refer), the first unit, *the question*, referentially points forward to the *What*-clause. There is thus a referential relationship between the two units.

and (c) an apposition requiring an obligatory marker apposition, such as *including*:

(4) *About 40 representatives of Scottish bodies*, including *the parents of some of the children flown to Corsica*, were addressed by an English surgeon and Dr. and by M. Naessens. (Survey of English Usage Corpus W.12.1–40)

In this example, the marker *including* is obligatory because it indicates that the reference of the noun phrase it precedes is included in the reference of the noun phrase in the first unit, resulting in the referential relationship of "inclusion" between the two units.

Because of their overall frequency, nominal appositions form the class of prototypical appositions. While the other three types of constructions that have been classified as appositions are more diverse in form and function, they are still within the class of appositions and are related through notions such as the two listed below (Lakoff 1987: 12):

centrality: this category contains members that "may be 'better examples' of that category than others."

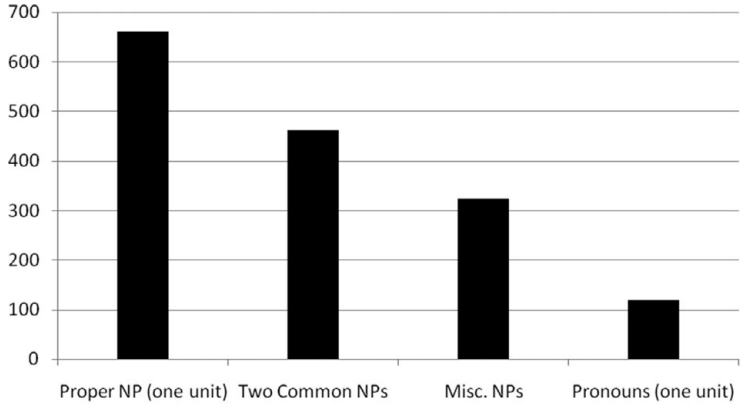

Figure 1.5 Forms of nominal appositions (raw frequencies)

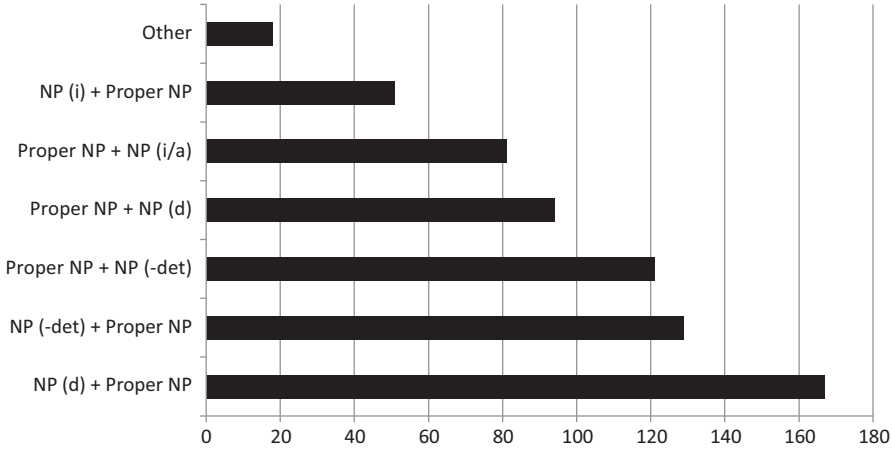

Figure 1.6 The form of appositions with proper nouns in one unit (raw frequencies)

membership gradience: some members have "degrees of membership and no clear boundaries."

Within the category of nominal appositions, Meyer (1992: 12) lists four different forms.

As Figure 1.5 demonstrates, APNs were the most frequent type of nominal apposition. Figure 1.6 lists the form and frequency of the various appositions containing a proper noun in one of the units (Meyer 1992: 12).

Each of the six forms (except for "other") are listed in the examples below, which contain identical content words:

(5)

(a)	NP (i) + Proper NP	a famous actress, Lauren Bacall
(b)	Proper NP + NP (i/a)	Lauren Bacall, a famous actress
(c)	Proper NP + NP (d)	Lauren Bacall, the famous actress
(d)	Proper NP + NP (-det)	Lauren Bacall, famous actress
(e)	NP (-det) + Proper NP	famous actress Lauren Bacall
(f)	NP (d) + Proper NP	the famous actress, Lauren Bacall

i = indefinite NP, i/a = indefinite NP, with attributive reference, d = definite NP, -det = NP lacking a determiner.

The examples given in (5) illustrate that the five forms are very closely related. And indeed, if their usage is examined in context, it is possible to see the systematic alternation of the forms. For instance, in the short excerpt below (taken from a single newspaper article), six APNs occur in three of the above five forms. Examples (6)(a) and (c) contain a proper noun followed by an indefinite noun phrase in an attributive relationship with the proper noun. Examples (6b and e) contain a proper noun followed by a noun phrase lacking a determiner. And examples (6)(d) and (f) contain a noun phrase lacking a determiner followed by a proper noun.

(6)

(a) ***…Thorneycroft, a pioneer in graffiti-removal patrols***, started cleaning her neighborhood two years ago. A retired accountant from Nabisco Brands Inc., Thorneycroft patrols her neighborhood at least twice a month. Her efforts have not gone unnoticed. This month she will receive two civic awards recognizing her relentless effort to address the graffiti problem.

(b) ***Mick Chase, board member for the Hawthorne Business Association***, said his group works closely with Thorneycroft to encourage businesses to clean graffiti off their buildings. "I think our businesses have a certain pride of ownership and don't like to see the graffiti around here," he said.

(c) ***John MacLeod, 56, a retired financial analyst for Bonneville Power Administration***, got tired of seeing profanities written on a new overpass near his home in the Parkrose area. He now covers the graffiti with paint supplied by the state highway division. The Southeast Uplift Neighborhood Program operates on donations and a $2,600 grant from the Oregon Community Foundation. The grant, received in the 1989–90 fiscal year, bought supplies for residents to use to clean up graffiti, said

(d) ***crime prevention organizer Peg Caliendo***. The group put together a brochure telling people how to organize cleaning programs and deal with graffiti.

(e) ***Helen Cheeks, crime prevention coordinator for Southeast Uplift***, said Tri-Met has given the group blanket permission to

clean up graffiti on its property, but the Postal Service hasn't. TriMet suggests groups or individuals should check before taking it on themselves to clean up the graffiti. The Postal Service, meanwhile, is trying to smooth the way for individuals and organizations to clean postal boxes. "We welcome the assistance in removing graffiti, but the details of the policy haven't been ironed out yet," said

(f) ***postal spokesman Bob Groff***. He said details of a new program, Adopt a Box, should be completed soon (ICE USA W2C-002b)

In context, although each of the examples would have differences in focus and meaning, they nevertheless are very closely interrelated. Within construction grammar, APNs can be viewed as a type of idiom, specifically as two of the types of idioms described in Fillmore, Kay, and O'Connor (1988). First of all, they have resemblances to what Fillmore, Kay, and O'Connor (1988: 505) label formal (i.e. schematic) idioms: "syntactic patterns dedicated to semantic and pragmatic purposes not knowable from their form alone." Schematic idioms differ from substantive idioms in that while their form is idiomatic, the lexical items that they contain can vary. For instance, one example of a schematic idiom that Fillmore, Kay, and O'Connor (1988) describe contains the pattern *the X-er the Y-er*. This pattern yields examples such as *the longer you work, the quicker you'll finish* or *the more revision you do, the more polished your essay will become*. In each of these examples, a grammatical pattern is filled with different lexical items, with only some parts of the pattern (the morpheme *–er* or the word *more*) being held constant. Although appositions of the form *Geoffrey Plimpton, police commissioner* contain no repeated lexical items from example to example, they are so frequent and vary so consistently that they warrant being classified as schematic idioms.

Moreover, APNs can also fit into a second class of idioms that Fillmore, Kay, and O'Connor (1988: 506) propose: "idioms with or without pragmatic point"; that is, idioms confined to a particular context or that occur in many different contexts. As Figure 1.7 illustrates, APNs predominated in press reportage (Meyer 1992: 117) and occurred much less frequently in fiction, learned writing, and casual conversation. These distributions exist because APNs, as will be shown, are pragmatically well-suited to press reportage, and when they occur outside this context, certain kinds of them, particularly those lacking an article in the first unit (e.g. *child television actor Jerry Mathers*) are pragmatically quite peculiar.

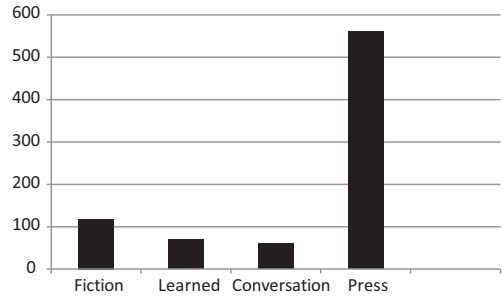

Figure 1.7 The distribution of APNs across registers (raw frequencies)

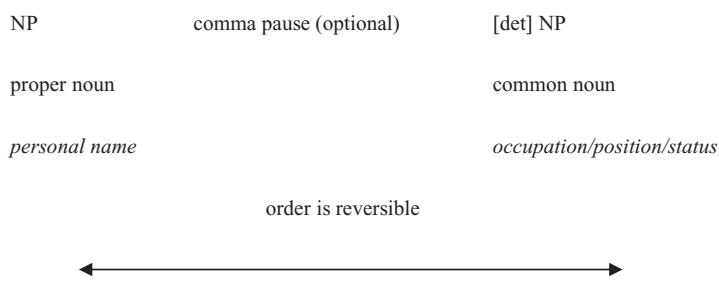

Figure 1.8 APN pattern

1.6.2.1 The Structure of APNs

The structure of the APN construction is described in the pattern in Figure 1.8:

This pattern describes the prototypical apposition containing one unit that is a proper noun. The examples in (7) illustrate all permutations of the pattern:

(7) (a) Barack Obama, a former president of the United States
(b) Barack Obama, former president of the United States
(c) The former president of the United States, Barack Obama
(d) Former president of the United States Barack Obama

Example (7)(a) illustrates the pattern of proper noun/comma pause/common noun (occupation). (7)(b) is similar except that the article is missing before the common noun. (7)(c) shows reversal of the two units. However, (7)(d) is problematic because there are constraints on the length of the first unit of the construction, largely because of the lack of a comma pause (intonation break) between the two units. As

Bell (1988: 332) notes, "In general, the greater the amount of structure in the descriptive NP [i.e. the first unit], the less acceptable determiner deletion becomes."

However, the length of the first unit of a pseudo-title varies from one variety of English to another. Meyer (2002: 162) found differences in the length of the first unit of a pseudo-title in the five components of the International Corpus of English he studied. In British and American English, relatively few first units exceeded four words in length:

(8) *Tory leader* Cllr Bob Blackman (ICE-GB:W2C-009 #54:3)

However, there was a statistically significant tendency for the first units (highlighted in the following) to be five words or longer in New Zealand and Philippine English:

(9) *Salamat and Presidential Adviser on Flagship Projects in Mindanao* Robert Aventajado (ICE-Philippines)

(10) *Oil and Gas planning and development manager* Roger O'Brien (ICE-NZ)

(11) *Time Magazine Asia bureau chief* Sandra Burton (ICE-Philippines)

(12) *Corporate planning and public affairs executive director* Graeme Wilson (ICE-NZ)

(13) *Autonomous Region of Muslime Mindanao police chief* Damming Unga (ICE-Philippines)

Thus, the actual length of the first unit in a pseudo-title is determined by both intonational limits on the length of the first unit as well as editorial practices on length dictated by the editorial practices of a newspaper that differ regionally.

1.6.3 The Communicative Functions of APNs

The form of APNs is very well suited to the genre – press reportage – in which they predominantly occur. Consequently, in other contexts, they sound rather odd. For instance, you might say to someone in a casual conversation:

(14) Jack Smith is a distinguished linguist.

However, you probably would not say:

(15) A distinguished linguist, Jack Smith, is having a drink with me later.

And you definitely would not say:

(16) Distinguished linguist Jack Smith is having a drink with me later.

Examples (17)–(18) illustrate the usefulness of APNs in a news story:

(17) Jessica Seinfeld's broccoli-spiked chicken nuggets recipes are all hers, a federal judge ruled Thursday. **Ms. Seinfeld**. . .did not copy from another author in her cookbook about sneaking vegetables into children's food, the judge said when she threw out a copyright infringement case brought by a competing author, Missy Chase Lapine. (NY Times, Sept. 11, 2009)

(18) **Ms. Seinfeld, the wife of the comedian Jerry Seinfeld**, did not copy from another author in her cookbook about sneaking vegetables into children's food. . .

In example (17), which contains the first sentence in the article from which the example was taken, the name *Jessica Seinfeld* is introduced into the text. In the next sentence, her name is mentioned again. However, unless you know who Ms. Seinfeld is, the significance of the story will not be fully understood by the reader. This is why in example (18), which contains the full sentence from the published news story, the second unit of an apposition is included: *the wife of the comedian Jerry Seinfeld*. In other words, it is not just newsworthy that Jessica Seinfeld was found not to have engaged in copyright infringement, but that she is also the wife of a famous comedian, Jerry Seinfeld.

News stories very frequently describe events that have taken place and the participants who are involved in the events. As a consequence, the names of the participants in news events play a key role in news stories. Langacker (2008: 259) notes the ability of nominals and finite clauses to create "a basic connection between the interlocutors and the content evoked." And, he continues, they do not simply refer to individuals but "evoke substantial bodies of information. . .widely shared within a speech community" (Langacker 2008: 316). He gives the example of George Washington, who is associated with certain commonly known attributes: general, president, honest, etc. However, in news stories, there is frequently no commonly shared knowledge of the people introduced into stories. The main function of APNs, then, is to provide this missing information.

1.7 Conclusions

This chapter described the role that corpus linguistics has played in the study of language from the past to the present, from pre-

electronic corpora to the many differing electronic corpora that currently exist. These corpora can be used to conduct linguistic research – both theoretical and applied – in many different areas, from genre variation (e.g. how language usage varies in spontaneous conversations versus public speeches) to research that is more applied and pedagogical (e.g. how the study of learner corpora can lead to insights for the teaching of English to individuals learning English as a second or foreign language).

Finally, a case study was presented to demonstrate that corpus analyses and various linguistic theories go hand in hand, and that such studies can do more than simply provide examples of constructions and document their frequency of occurrence. If this is the only information that a corpus analysis could provide, then corpus linguistics would have at best a marginal role in the field of linguistics. Instead, linguistic theories can be used to explain why particular frequencies and examples actually exist in a corpus: in other words, to discuss how theory and data interact. This is why corpus linguistics has grown as an area of linguistics and why many people now are using linguistic corpora for many different purposes.

2 Planning the Construction of a Corpus

The first step in building a corpus is to decide what the ultimate purpose of the corpus will be. This decision is important because it will determine exactly what the corpus will look like: what types of texts will be included in it (e.g. spoken, written, both), how long the individual texts will be (e.g. full length, text excerpts), how detailed the annotation will need to be (e.g. word-class tagging or a purely lexical corpus with minimal annotation), and so forth. For instance, if the corpus is to be used primarily for grammatical analysis (e.g. the analysis of relative clauses or the structure of noun phrases), the corpus can consist simply of text excerpts rather than complete texts, and will minimally need part-of-speech tags. On the other hand, if the corpus is intended to permit the study of discourse features, then it will have to contain lengthier texts, and some system of tagging to describe the various features of discourse structure.

To explore the process of planning a corpus, this chapter first provides an overview of the methodological assumptions that guided the compilation of two general-purpose corpora – the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA) – and two more specialized corpora – the Corpus of Early English Correspondence (CEEC) and the International Corpus of Learner English (ICLE). The remainder of the chapter then focuses more specifically on the individual methodological considerations (e.g. ensuring that a corpus is "balanced") that anyone planning to create a corpus needs to address.

2.1 The British National Corpus

At approximately 100 million words in length, the British National Corpus (BNC) (see Table 2.1) is one of the larger and more carefully planned corpora ever created. Most of the corpus (about 90 percent) consists of various types of written British English, with the

Table 2.1 *The composition of the British National Corpus (adapted from www.natcorp.ox.ac.uk/docs/URG/BNCdes.html)*

Speech			
Type	*# of Texts*	*# of Words*	*% of Spoken Corpus*
Demographically Sampled	153	4,233,955	41
Educational/ Informative	169	1,646,380	16
Business	129	1,282,416	12
Public/ Institutional	262	1,672,658	16
Leisure	195	1,574,442	15
Total	908	10,409,851	
Writing			
Type	*# of Texts*	*# of Words*	*% of Written Corpus*
Imaginative	476	16,496,420	19
Informative: Natural & pure science	146	3,821,902	4
Informative: Applied science	370	7,174,152	8
informative: Social science	526	14,025,534	16
Informative: World affairs	483	17,244,534	20
Informative: Commerce & finance	295	7,341,163	8
Informative: Arts	261	6,574,857	7
Informative: Belief & thought	146	3,037,533	3
Informative: Leisure	438	12,237,834	
Total	2,703	87,953,929	99

remainder (about 10 percent) comprised of different types of spoken British English. Even though writing dominates in the BNC, the amount of spoken language in the corpus is especially valuable, largely because it represents types of speech, such as spontaneous conversations, that are not widely found in other corpora.

In planning the collection of texts for the BNC, a number of decisions were made beforehand:

(1) Texts included in the corpus were selected according to domain (the categories listed in Table 2.1), time (a range of dates), and medium (book, periodical, etc.).
(2) Even though the corpus was planned to contain both speech and writing, most of the corpus (90 percent) ultimately consisted of written texts and the remainder (10 percent) spoken texts. These distributions were largely a consequence of the amount of time and effort it would take to record and transcribe speech.
(3) A variety of different genres of speech and writing would be gathered for inclusion in the corpus in order to ensure representation of the many different genres of speech and writing that exist. However, greater amounts of some genres were collected than others. For instance, because imaginative writing (e.g. fiction) is more common than other types of writing, more samples of this type were collected than other types. Of course, this principle could not be followed consistently in the corpus: Spontaneous conversations are the most common form of human communication, but as noted above, their collection and transcription are too laborious a process to include more of this type of speech in a corpus.
(4) Each genre would be divided into text samples, and each sample would not exceed 45,000 words in length. Thus, the goal was to represent as many different writers and speakers as possible – a goal that unfortunately would prevent the study of certain kinds of discourse features that would be present in texts that are of full length.
(5) A number of variables would be controlled for in the entire corpus, such as the age and gender of speakers and writers.
(6) For the written part of the corpus, a careful record of a variety of variables would be kept, including when and where the texts were written or published; whether they were books, articles, or manuscripts; who their target audience was; and the age and gender of individuals comprising the target audience.
(7) For the demographically sampled spoken samples, texts would be collected from individuals representing the major dialect regions of Great Britain and the various social classes existing within these regions.

Currently, there is a newer version of the BNC, BNC2014, that is under development and that will replicate the structure of the original corpus.

2.2 The Corpus of Contemporary American English

Like the BNC, the Corpus of Contemporary American English (COCA) contains various samples of different kinds of

Table 2.2 *The Corpus of Contemporary American English (as of April 7, 2021)(www.english-corpora.org/coca/)*

Type	*# of Words*
Speech	
Transcripts of unscripted television and radio programs	127,396,916
TV/Movies subtitles	128,013,334
Written	
Fiction	119,505,014
Magazines	127,352,014
Newspapers	122,959,393
Academic journals	120,988,348
Blogs	125,496,215
Web	129,899,426
Total	1,001,610,938

spoken and written English, with the exception that the corpus represents American rather than British English (see Table 2.2).

In addition, COCA is a monitor corpus; that is, a corpus to which new texts are added on an ongoing basis. Moreover, COCA is also much lengthier than the BNC: Currently (as of April 2021), it is 1 billion words in length. While the two corpora share several commonalities, they differ in many key areas:

(1) While both COCA and the BNC each contain samples of speech, COCA contains a more restricted sampling of spoken English; only shows broadcast on television or radio as well as TV and movie subtitles. Unlike the BNC, it contains no spontaneous conversations of individuals having, for example, a casual conversation during breakfast, or a discussion of a current movie that a group of individuals had seen.
(2) The samples of speech included in the BNC were transcribed from recordings of all the spoken samples included in the corpus. In contrast, the samples of speech included in COCA were based on transcripts created by a separate party (e.g. the broadcast network from which the sample was taken). The advantage of using broadcast transcripts is that it is not necessary to spend considerable time manually transcribing recorded speech. Consequently, a corpus as large as COCA can be created in a relatively short period of time. But one of the potential disadvantages of using broadcast transcriptions is that their accuracy cannot be guaranteed, a point discussed in greater detail later in the chapter.

(3) The BNC contains numerous texts that are subject to copyright restrictions. So that users can make full use of these texts, creators of the BNC received permission from the copyright holders allowing users of the corpus to make use of the copyrighted material in their research (with certain restrictions; cf. www.natcorp.ox.ac.uk/docs/licence.pdf). Likewise, COCA contains a significant amount of copyrighted material. Therefore, it can be used at no cost by accessing the web interface, and search results will contain a limited amount of text (and context) in sentences containing words or phrases being searched for. However, the corpus can be purchased for a fee, allowing full use of the complete corpus, with certain restrictions: at the juncture of 200 words in a given text, 10 words are elided and replaced with 10 ampersands: @ (www.corpusdata.org/limitations.asp).

2.3 The Corpus of Early English Correspondence (CEEC)

The CEEC is actually a group of historical corpora that were created over the years at the University of Helsinki.

Two of the corpora in Table 2.3 are available for use by the research community. One corpus (CEECS) is a sampler corpus; that is, it contains a subset of texts taken from CEEC, minus texts that are subject to copyright restrictions. A second corpus (PCEEC) contains most of the texts in CEEC that have been tagged and parsed: Part-of-speech tags have been assigned to each word, and various grammatical categories, such as noun phrases and verb phrases, have been assigned grammatical tags. Moreover, CEECE is similar in content to CEEC, except that it contains correspondence beyond the time frame (1681) when texts in CEEC conclude. The CEECSU covers the same

Table 2.3 *The various versions of the Corpus of Early English Correspondence (https://varieng.helsinki.fi/CoRD/corpora/CEEC/index.html)*

Corpus	Time Covered	Words	Letters	Writers	Collections
CEEC	1410?–1681	2.7 million	6039	778	96
CEECS	1418–1680	0.45 million	1147	194	23
PCEEC	1410?–1681	c. 2.2 million	4979	657	84
CEECE	1681–1800	c. 2.2 million	c. 4900	>300	74
CEECSU	1402–1663	c. 0.44 million	c. 900	>100	20

time period as CEEC but provides a greater range of writers, particularly women, than CEEC.

All of the corpora in Table 2.3 have a number of design and compilation features that distinguish them from the BNC and COCA:

(1) The corpora were designed "to test the applicability of sociolinguistic methods to historical data" (www.helsinki.fi/varieng/CoRD/corpora/CEEC/compilation.html). Because no spoken data exists from this period, letters were chosen because, unlike other types of historical documents, they are best suited to study how various sociolinguistic variables, such as age, social class, or gender, affect language use. Other types of written text (e.g. records from trials, depositions of witnesses) can also be used to study more speech-like language in other contexts. The Old Bailey Corpus (version 2.0) contains 24.4 million words of speech from trials appearing in the Proceedings of the Old Bailey between 1720–1913 (www.clarin.eu/showcase/old-bailey-corpus-20-1720-1913).
(2) But while gender balance is more easily achieved in modern corpora such as the BNC or COCA, in CEEC it was much more difficult, largely because literacy rates were much higher among men during this period than women. Consequently, fewer female writers were included in the corpus than male writers.
(3) Most of the letters were digitized by scanning texts from edited books in which they appeared. However, some of the texts had to be typed in manually because they did not appear in edited editions, a very "laborious method of keying in the text" (www.helsinki.fi/varieng/CoRD/corpora/CEEC/generalintro.html).

2.4 The International Corpus of Learner English (ICLE)

The ICLE (version 2) is a learner corpus: a type of corpus containing samples of speech or writing produced by individuals learning English as a foreign or second language. The ICLE is restricted to writing and contains samples of written English produced by individuals at an advanced level of proficiency learning written English as a foreign language. The samples of written English included in the corpus were obtained from native speakers of 12 European languages and 4 non-European languages:

Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Tswana, Turkish

Some of the design principles of ICLE are similar to those discussed in previous sections with the differences mainly relating to the unique features of learner corpora:

(1) At 3,753,030 words in length, ICLE is smaller than either the BNC or COCA. But like these two corpora, it is divided into samples, ranging in length from 500 to 1,000 words. However, the type of writing included in the corpus is restricted to primarily argumentative writing (Granger et al. 2009: 5).
(2) Detailed ethnographic information was recorded for each individual whose writing was included in the corpus, including such features as age, gender, mother tongue, the region in which the writer lives (for languages which are spoken in more than one country), and the other languages the writer may speak (Granger et al. 2009: 4).
(3) Like the BNC and COCA, ICLE is lexically tagged and comes with a concordancing program that has been customized to search for specific tags or combinations of tags included in the corpus and to also study effects on usage such as the age, gender, mother tongue, and so forth of the writers whose texts have been included in the corpus.

The preceding sections provided a brief overview of four different corpora and some of the methodological considerations that guided the development of these corpora. The remainder of the chapter explores these considerations in greater detail, focusing on such topics as the factors determining, for instance, how lengthy a corpus should be, what kinds of texts should be included in a corpus, and other issues relevant to the design of linguistic corpora.

2.5 What Is a Corpus?

The corpora discussed so far have a clearly identifiable structure: we know how lengthy they are, what texts are in them, and which genres the texts represent. Recently, however, the Web itself has increasingly been used as a source of data for linguistic analysis, giving rise to the notion of "the web as corpus." But as Gatto (2014: 35) observes, the idea of the web as a corpus runs counter to long-held conceptions of exactly what a corpus is:

> On the one hand, the traditional notion of a linguistic corpus as a body of texts rests on some correlate issues, such as finite size, balance, part-whole relationship and permanence; on the other hand, the very idea of a web of

texts brings about notions of non-finiteness, flexibility, decentering/re-centering, and provisionality.

With a corpus such as the BNC, we know precisely what kinds and types of English are being analyzed. With web-based material, however, no such certainty exists.

To determine the kinds of texts that exist on the Web, Biber, Egbert, and Davies (2015) conducted an empirical study based on a collection of texts taken from the Corpus of Global Web-Based English (GloWbE), a corpus that is 1.9 billion words in length and that contains samples of English from 20 different countries in which English is used. Biber, Egbert, and Davies (2015) developed a very carefully planned methodology for extracting a representative body of texts in the corpus from the Web, and then trained a group of evaluators to categorize the texts into specific registers.

They found that three registers predominated: narrative (31.2 percent), informational description/explanation (14.5 percent), and opinion (11.2 percent). In the narrative category, general news reports (52.5 percent) and sports reports (16 percent) were most frequent (p. 24). The informational description/explanation category contained, for instance, research articles and technical reports, though interestingly "academic research articles – the focus of an extensive body of corpus-based research – comprise less than 3 percent of the general "'informational' register" (p. 26). This register also led to considerable disagreement among evaluators, with 53.9 percent of the documents receiving "no majority agreement" as to which register in which they should be classified. Within the opinion register, 37.9 percent of the texts were blogs, and another 21 percent were reviews (p. 28).

Biber, Egbert, and Davies (2015) also discovered a number of texts that were hybrid in nature: "documents that combine multiple communicative purposes in a single text" (p. 31). A full 29.2 percent of documents analyzed in this study were classified into this category (p. 23), raising interesting questions about not just web-based corpora but the system of classification of registers in general.

Future chapters will explore in greater detail corpus analyses of web data, and how tools such as the WebCorp concordancing program (www.webcorp.org.uk/live/) can be used to extract data directly from the Web. But the remaining sections in this chapter will focus primarily on issues related to the construction of traditional corpora.

2.6 Corpus Size

The Brown Corpus, released in 1962, is one million words in length. A more recent corpus, the Corpus of Web-Based Global English, is 1.9 billion words in length. Although there are many reasons for the disparity in the length of these corpora, the primary reasons can be attributed to advances in technology and the easy availability of texts in electronic form.

First generation corpora, such as Brown and LOB, were relatively short (each of one million words in length), mainly because of the logistical difficulties that computerizing a corpus created. For instance, all of the written texts for the Brown Corpus had to be keyed in by hand, a process requiring a tremendous amount of very tedious and time-consuming typing. In contrast, those creating second generation corpora have had the advantage of numerous technological advances that have automated the process of corpus creation. For instance, the availability of optical scanners made it easier to convert printed texts into digital formats. More recently, so many texts are available in electronic formats that with very little work, they can easily be adapted for inclusion in a corpus. As a consequence, second generation corpora are regularly 100 million words in length or even longer.

But while written texts can easily be digitized, technology has not progressed to the point where it can greatly expedite the collection and transcription of speech, especially spontaneous conversations: There is much work involved in recording this type of speech and manually transcribing it. Speech recognition software, which converts speech into writing, has been greatly improved over the years, but it works best with a single speaker who has trained the software to recognize his/her speech characteristics. These logistical realities explain why 90 percent of the BNC (a second-generation corpus) contains written texts and only 10 percent spoken texts.

Other corpora, such as COCA, which contain a significant amount of speech, generally consist of spoken samples, such as broadcast discussions, for which commercially produced transcripts are available. But as will be discussed in a later section, the wide availability of transcription services has made the transcription of speech more financially feasible. In addition, the creation of the spoken BNC2014 and the London-Lund Corpus demonstrate the feasibility of creating corpora with significant amounts of spoken language.

Ultimately, the length of a corpus is best determined by its intended use. Davies (2015: 12–18) provides a useful guide for determining how lengthy a corpus needs to be to accurately describe particular linguistic structures. Davies analyzed three different corpora of varying length: the Brown Corpus (one million words), the BNC (100 million words), and COCA (at that time 500 million+ words). His goal was to determine the extent to which the length of a corpus could provide valid information on 10 different linguistic constructions, including individual lexical items; frequently occurring grammatical structures, such as modal verbs and passives; collocations, and an assortment of other commonly studied grammatical items.

Not surprisingly, Davies found that individual lexical items were better studied in larger corpora than in shorter corpora. For instance, while adjectives such as *fun* or *tender* are among the group of adjectives that are most common in COCA, in the Brown Corpus, they occurred five times or less. In contrast, certain types of syntactic structures, such as modal verbs, have more even distributions across the three corpora, thus being one of the few areas "where Brown provides sufficient data" (Davies 2015: 15). Overall, COCA provides many more examples of the 10 types of constructions that Davies studied. However, in some instances, the number of occurrences is so high that even though the BNC may contain lower frequencies, the numbers are still high enough to permit valid studies. For instance, while COCA contains 2,900,000 *be* passives, the BNC contains 890,000 examples (Davies 2015: 15), a number of occurrences that is certainly sufficient enough to study this type of passive. Moreover, frequencies alone can only be suggestive, since different types of studies may concentrate on texts with lower frequencies, especially if such studies are more qualitative in focus.

Biber (1993) provides a different mechanism for estimating the necessary size of a corpus for the study of particular linguistic constructions. His approach employs statistical formulas that take the frequency with which linguistic constructions are likely to occur in a corpus and then calculate how large the corpus will have to be to validly study the distribution of the constructions. In this study (based on information in Biber 1988), Biber (1993: 253–4) examined 481 text samples that occurred in 23 different genres of speech and writing. He found that reliable information could be obtained on frequently occurring linguistic items such as nouns in as few as 59.8 text samples. On the other hand, infrequently occurring grammatical constructions such as conditional clauses required a much

larger number of text samples (1,190) for valid information to be obtained. Biber (1993: 254) concludes that the "most conservative approach" would be to base the ultimate size of a corpus on "the most widely varying feature"; that is, those linguistic constructions, such as conditional clauses, that require the largest sample size for reliable studies to be conducted. A corpus of 1,190 samples would therefore be 2,380,000 words in length (if text samples were 2,000 words in length, the standard length in many first-generation corpora).

Unfortunately, such calculations presuppose that one knows precisely what linguistic constructions will be studied in a corpus. The ultimate length of a corpus might therefore be better determined not by focusing too intently on the overall length of the corpus but by focusing more on the internal structure of the corpus: the range of genres one wishes to include in the corpus, the length and number of individual text samples required to validly represent the genres that will make up the corpus, and the demographic characteristics of the individuals whose speech and writing will be chosen for inclusion in the corpus.

2.7 The Internal Structure of a Corpus

The BNC, as Table 2.1 indicates, contains a diverse range of spoken and written genres. A plurality of the spoken texts (41 percent) was "demographically sampled"; that is, they consisted of a variety of spontaneous dialogues and monologues recorded from individuals living in various parts of Great Britain. The remainder of the spoken samples contained a fairly even sampling (12–14 percent) of monologues and dialogues organized by purpose. For instance, those texts in the genre of education included not just classroom dialogues and tutorials but lectures and news commentaries as well (Crowdy 1993: 263). A plurality of the written texts (22 percent) was "imaginative"; that is, they represented various kinds of fictional and creative writing. Slightly less frequent were samples of writing from world affairs (18 percent), the social sciences (15 percent), and several other written genres, such as the arts (8 percent) and the natural sciences (4 percent).

If the BNC is compared with the International Corpus of English (ICE), a collection of comparable corpora representing the major varieties of English spoken worldwide, it turns out that while the two corpora contain the same range of genres, the genres are much more specifically delineated in ICE Corpora (see Table 2.4) than they

Table 2.4 *The composition of the International Corpus of English (adapted from Greenbaum 1996a: 29–30)*

Type	*Number of Texts*	*Length*	*% of Spoken Corpus*
Speech			
Dialogues	180	360,000	60
Private	100	200,000	33
-direct conversations	90	180,000	30
-distanced conversations	10	20,000	3
Public	80	160,000	27
-class lessons	20	40,000	7
-broadcast discussions	20	40,000	7
-broadcast interviews	10	20,000	3
-parliamentary debates	10	20,000	3
-business transactions	10	20,000	3
Monologues	120	240,000	40
Unscripted	70	140,000	23
-spontaneous commentaries	20	40,000	7
-speeches	30		
-demonstrations	10	20,000	3
-legal presentations	10	20,000	3
Scripted	50	100,000	17
-broadcast news	20	40,000	7
-broadcast talks	20	40,000	7
-speeches (not broadcast)	10	20,000	3
Total	300	600,000	99
Writing Type	***Number of Texts***	***Length***	***% of Written Corpus***
Nonprinted	50	100,000	25
-student untimed essays	10	20,000	5
-student examination essays	10	20,000	5
-social letters	15	30,000	8
-business letters	15	30,000	8
Printed	150	300,000	75
-informational (learned):	40	80,000	20
-informational (popular)	40	80,000	20
-informational (reportage)	20	40,000	10
-instructional:	20	40,000	10
-persuasive (press editorials)	10	20,000	5
-creative (novels, stories)	20	40,000	10
Total	200	400,000	101

are in the BNC. For instance, in both corpora, 60 percent of the spoken texts are dialogues and 40 percent are monologues. In the ICE, this division is clearly reflected in the types of genres making up the spoken part of the corpus. For instance, in the category of speech, there are dialogues and monologues. Dialogues can be either private (e.g. direct conversations) or public (e.g. broadcast discussions). In the BNC, on the other hand, dialogues and monologues are interspersed among the various genres (e.g. business, leisure) making up the spoken part of the corpus (Crowdy 1993: 263).

Likewise, the other spoken genres in the ICE (e.g. scripted and unscripted speeches) are included in the BNC but within the general areas outlined in Table 2.1. In both corpora, there is a clear bias towards spontaneous dialogues: In the ICE, 33 percent of the spoken texts consist of direct conversations or telephone calls; in the BNC, 41 percent of the spoken texts are of this type (although some of the texts in the category are spontaneous monologues as well).

While the amount of writing in the BNC greatly exceeded the amount of speech, just the opposite is true in the ICE Corpus: only 40 percent of the texts are written. While creative (or imaginative) writing was the most common type of writing in the BNC, in the ICE it is not as prominent. More prominent were learned and popular examples of informational writing: writing from the humanities, social and natural sciences, and technology (40 percent of the written texts). These categories are also represented in the BNC, although the BNC makes a distinction between the natural, applied, and social sciences and, unlike the ICE, does not include equal numbers of texts in each of these categories. The ICE also contains a fairly significant number (25 percent) of nonprinted written genres (such as letters and student writing), while only 5–10 percent of the BNC contains these types of texts.

To summarize, while there are differences in the composition of the ICE and BNC, overall the two corpora represent similar genres of spoken and written English. The selection of these genres raises an important methodological question: why these genres and not others?

The answer is that both the ICE and BNC are *multi-purpose corpora*; that is, they are intended to be used for a variety of different purposes, ranging from studies of vocabulary, to studies of the differences between various national varieties of English, to studies whose focus is grammatical analysis, to comparisons of the various genres of English. For this reason, each of these corpora contains a broad range of genres. But in striving for breadth of coverage, some compromises

had to be made in each corpus. For instance, while the spoken part of the ICE contains legal cross-examinations and legal presentations, the written part of the corpus contains no written legal English. Legal written English was excluded from the ICE on the grounds that it is a highly specialized type of English firmly grounded in a tradition dating back hundreds of years, and thus does not truly represent English as it is written in the 1990s (the years during which texts for the ICE are being collected). The ICE also contains two types of newspaper English: press reportage and press editorials. However, as Biber (1988: 180–96) notes, newspaper English is a diverse genre, containing not just reportage and editorials but, for instance, feature writing as well – a type of newspaper English not represented in the ICE.

While both the ICE and the BNC have a very clearly defined internal structure, the Corpus of Contemporary American English (COCA) differs somewhat in that it contains collections of texts (of varying length) representing five major registers: spoken (transcripts of dialogical speech taken from various television and radio shows), fiction, newspapers, magazines, and academic writing. Each of these registers contains 103–110 million words of text collected during the years 1990 through 2019 with each year containing approximately 20 million words of text (http://corpus.byu.edu/coca/?f=texts_e). Because of the overall size of the corpus (one billion words), each of the registers is more fully represented than those occurring in either the ICE or BNC. Additionally, COCA has a diachronic component because it contains text collected over a span of 19 years. While registers within the ICE and BNC are divided into samples no lengthier than 45,000 words, each register in COCA contains samples of varying and indeterminate length.

As will be discussed in subsequent chapters, *general-purpose corpora* are useful for the analysis of many kinds of grammatical constructions. However, for those wishing to study a particular register, say press reportage, these corpora do not contain enough samples of the individual genres to permit full-scale studies. Consequently, many *special-purpose corpora* have been developed. These are corpora with a more specific focus. Two such corpora were discussed earlier in the chapter: the CEEC and the ICLE. But there are many additional such corpora as well. For instance, the Michigan Corpus of Academic Spoken English (MICASE) was created to study the type of speech used by individuals conversing in an academic setting: class lectures, class discussions, student presentations, tutoring sessions, dissertation

defenses, and many other kinds of academic speech (Powell and Simpson 2001: 34–40). As of March 4, 2022, 152 samples (1,848,364 words) could be searched on the MICASE web site: https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase.

At the opposite spectrum of special-purpose corpora like MICASE are those which have specialized uses but not for genre studies. Treebank-3 consists of a heterogeneous collection of texts totaling 100 million words, including one million words of text taken from the 1989 *Wall Street Journal*, as well as tagged and parsed versions of the Brown Corpus. The reason that a carefully selected range of genres is not important in this corpus is that the corpus is not intended to permit genre studies but to, for instance, "train" taggers and parsers to analyze English: to present them with a sufficient amount of data so that they can "learn" the structure of numerous constructions and thus produce a more accurate analysis of the parts of speech and syntactic structures present in English. And to accomplish this goal, all that is important is that a considerable number of texts be available, and less important are the genres from which the texts were taken.

Because of the wide availability of written and spoken material, it is relatively easy to collect material for modern-day corpora such as the BNC and ICE: The real work is in recording and transcribing spoken material, for instance, or obtaining copyright clearance for written material. With historical corpora, however, collecting texts from the various genres existing in earlier periods is a much more complicated undertaking.

In selecting genres for inclusion in historical corpora, the goals are similar to those for modern-day corpora: to find a range of genres representative of the types of English that existed during various historical periods of English. Consequently, there exist multi-purpose corpora, such as the Helsinki Corpus, which contains a range of different genres (sermons, travelogues, fiction, drama, etc.), as well as specialized corpora, such as the previously mentioned Corpus of Early English Correspondence, a corpus of letters written during the middle English period, with the original CEEC containing letters from late middle English and early modern English and a later version of the corpus (CEECE) letters from the eighteenth century. Other corpora, such as the Corpus of English Dialogues (1560–1760) and the Old Bailey Corpus (1674–1913), contain letters representing later periods.

In gathering material for corpora such as these, the corpus compiler must deal with the fact that many of the genres that existed in earlier

periods are either unavailable or difficult to find. For instance, even though spontaneous dialogues were as common and prominent in earlier periods as they are today, there were no tape recorders around to record speech. Therefore, there exists no direct record of speech. However, this does not mean that we cannot get at least a sense of what speech was like in earlier periods. In her study of early American English, Kytö (1991: 29) assembled a range of written texts that are second-hand representations of speech: various "verbatim reports," such as the proceedings from trials and meetings and depositions obtained from witnesses, and transcriptions of sermons. Of course, it can never be known with any great certainty exactly how close to the spoken word these kinds of written texts are. Nevertheless, texts of this type give us the closest approximation of the speech of earlier periods that we will ever be able to obtain.

In other situations, a given genre may exist but be underrepresented in a given period. In his analysis of personal pronouns across certain periods in the Helsinki Corpus, Rissanen (1992: 194) notes that certain genres were difficult to compare across periods because they were "poorly represented" during certain periods: histories in Old English and laws and correspondence in Middle English. Other types of genres exist in earlier periods but are defined differently than they are in the modern period. The Helsinki Corpus includes the genre of science in the Old English period, but the texts that are included focus only on astronomy, a reflection of the fact that science played a much more minimal role in medieval culture than it does today.

2.8 The Length of Individual Text Samples to Be Included in a Corpus

Corpora vary in terms of the length of the individual text samples that they contain. First generation corpora, such as the Brown or London-Oslo-Bergen (LOB) corpora, contain 2,000-word samples. An early corpus of spoken British English, the London-Lund Corpus, contains 5,000-word samples. In the Helsinki Corpus, text samples range from 2,000 to 10,000 words in length. And samples within the BNC vary in length, but are no longer than 40,000 words.

While many corpora contain only text samples, others contain entire texts. For instance, the Lampeter Corpus of Early Modern English Tracts, which is c. 1.1 million words in length, consists of complete texts ranging in length from 3,000 to 20,000 words. The

Corpus of Global Web-Based English (1.9 billion words) also contains complete texts of varying length.

Ideally, it would be desirable to include complete texts in corpora, since even if one is studying grammatical constructions, it is most natural to study these constructions within the context of a complete text rather than only part of that text. However, there are numerous logistical obstacles that make the inclusion of complete texts in corpora nearly impossible. For instance, many texts, such as books, are quite lengthy, and to include a complete text in a corpus would not only take up a large part of the corpus but require the corpus compiler to obtain permission to use not just a text excerpt, a common practice, but an entire text, a very uncommon practice. In general, it is quite difficult to obtain permission to use copyrighted material. To avoid copyright infringement, those using the BYU corpora (such as the Corpus of Contemporary American English) are only allowed to view "snippets" in search returns of grammatical items in the corpora they are studying.

Of course, just because only text samples are included in a corpus does not mean that sections of texts ought to be randomly selected for inclusion in a corpus. It is possible to take excerpts that themselves form a coherent unit. For instance, many parts of spoken texts form coherent units themselves, containing sections that have their own beginnings, middles, and ends. Likewise, for written texts, one can include the first 2,000 words of an article, which contains the introduction and part of the body of the article, or one can take the middle of an article, which contains a significant amount of text developing the main point made in the article, or even its end. Many samples in the ICE also consist of composite texts: a series of complete short texts that total 2,000 words in length. For instance, personal letters are often less than 2,000 words, and a text sample can be comprised of complete letters totaling 2,000 words. For both the spoken and written parts of the corpus, not all samples are exactly 2,000 words: a sample is not broken off in mid-sentence but at a point (often over or just under the 2,000-word limit) where a natural break occurs. But even though it is possible to include coherent text samples in a corpus, creators and users of corpora simply have to acknowledge that corpora are not always suitable for many types of discourse studies, and that those wishing to carry out such studies will simply have to assemble their own corpora for their own personal use.

In including short samples from many different texts, corpus compilers are assuming that it is better to include more texts from many

different speakers and writers than fewer texts from a smaller number of speakers and writers. And there is some evidence to suggest that this is the appropriate approach to take in creating a corpus. Biber (1990 and 1993: 248–52) conducted an experiment in which he used the LOB and London-Lund Corpora to determine whether text excerpts provide a valid representation of the structure of a particular genre. Biber divided the LOB and London-Lund Corpora into 1,000-word samples: he took two 1,000-word samples from each 2,000-word written sample of the LOB Corpus, and he divided each 5,000-word sample of speech from the London-Lund Corpus into four 1,000-word samples. In 110 of these 1,000-word samples, Biber calculated the frequency of a range of different linguistic items, such as nouns, prepositions, present and past tense verb forms, passives, and so forth. Biber concluded that 1,000-word excerpts are lengthy enough to provide valid and reliable information on the distribution of *frequently occurring* linguistic items. That is, if one studied the distribution of prepositions in the first 1,000 words of a newspaper article totaling 10,000 words, for instance, studying the distribution of prepositions in the entire article would not yield different distributions: the law of diminishing returns is reached after 1,000 words. On the other hand, Biber found that *infrequently occurring* linguistic items (such as relative clauses) cannot be reliably studied in a short excerpt; longer excerpts are required.

In addition to studying the distribution of word categories, such as nouns or prepositions, Biber (1993: 250) calculated the frequency with which new words are added to a sample as the number of words in the sample increases. He found, for instance, that humanities texts are more lexically diverse than technical prose texts (p. 252); that is, that as a humanities text progresses, there is a higher likelihood that new words will be added as the length of the text increases than there will be in a technical prose text. This is one reason that lexicographers need such large corpora to study vocabulary trends, since so much vocabulary (in particular, open-class items such as nouns and verbs) occurs so rarely. And as more text is considered, there is a greater chance (particularly in humanities texts) that new words will be encountered.

Biber's (1993) findings would seem to suggest that corpus compilers ought to greatly increase the length of text samples to permit the study of infrequently occurring grammatical constructions and vocabulary. However, Biber (1993: 252) concludes just the opposite, arguing that "Given a finite effort invested in developing a corpus, broader

linguistic representation can be achieved by focusing on diversity across texts and text types rather than by focusing on longer samples from within texts." In other words, corpus compilers should strive to include more different kinds of texts in corpora rather than lengthier text samples. Moreover, those using corpora to study infrequently occurring grammatical constructions will need to go beyond currently existing corpora and look at additional material on their own.

2.9 Determining the Number of Texts and Speakers and Writers to Include in a Corpus

Related to the issue of how long text samples should be in a corpus is precisely how many text samples are necessary to provide a representative sampling of a genre, and what types of individuals ought to be selected to supply the speech and writing used to represent a genre. These two issues can be approached from two perspectives: from a purely linguistic perspective, and from the perspective of sampling methodology, a methodology developed by social scientists to enable researchers to determine how many "elements" from a "population" need to be selected to provide a valid representation of the population being studied. For corpus linguists, this involves determining how many text samples need to be included in a corpus to ensure that valid generalizations can be made about the genre, and what range of individuals need to be selected so that the text samples included in a corpus provide a valid representation of the population supplying the texts.

There are linguistic factors that need to be considered in determining the number of samples of a genre to include in a corpus, considerations that are quite independent of general sampling issues. If the number of samples included in the various genres of the BNC and ICE Corpora are surveyed, it is immediately obvious that both of these corpora place a high value on spontaneous dialogues, and thus contain more samples of this type of speech than, say, scripted broadcast news reports. This bias is a simple reflection of the fact that those creating the BNC and ICE Corpora felt that spontaneous dialogues are a very important type of spoken English and should therefore be amply represented. The reason for this sentiment is obvious: while only a small segment of the speakers of English creates scripted broadcast news reports, all speakers of English engage in spontaneous dialogues.

Although it is quite easy to determine the relative importance of spontaneous dialogues in English, it is far more difficult to go through every potential genre to be included in a corpus and rank its relative importance and frequency to determine how much of the genre should be included in the corpus. And if one did take a purely "proportional" approach to creating a corpus, Biber (1993: 247) notes, the resultant corpus "might contain roughly 90 percent conversation and 3 percent letters and notes, with the remaining 7 percent divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writings" since these figures provide a rough estimate of the actual percentages of individuals that create texts in each of these genres. To determine how much of a given genre needs to be included in a corpus, Biber (1993) argues that it is more desirable to focus only on linguistic considerations, specifically how much internal variation there is in the genre. As Biber (1988: 170f.) has demonstrated in his pioneering work on genre studies, some genres have more internal variation than others, and consequently, more examples of these genres need to be included in a corpus to ensure that the genres are adequately represented. For instance, Biber (1988: 178) observes that even though the genres of official documents and academic prose contain many sub-genres, the sub-genres in official documents (e.g. government reports, business reports, and university documents) are much more linguistically similar than the sub-genres of academic prose (e.g. the natural and social sciences, medicine, and the humanities). That is, if the use of a construction such as the passive is investigated in the various sub-genres of official documents and academic prose, there will be less variation in the use of the passive in the official documents than in the academic prose. This means that a corpus containing official documents and academic prose will need to include more academic prose to adequately represent the amount of variation present in this genre.

In general, Biber's (1988) study indicates that a high rate of internal variation occurs not just in academic prose but in spontaneous conversation (even though there are no clearly delineated sub-genres in spontaneous conversation); spontaneous speeches; journalistic English (though Biber analyzed only press reportage and editorials); general fiction; and professional letters. Less internal variation occurs in official documents (as described above); science fiction; scripted speeches; and personal letters.

Because the BNC is a relatively lengthy corpus, it provides a sufficient number of samples of genres to enable generalizations to

Table 2.5 *Pseudo-titles in four corpora*

Time Period	British English	American English
1930s	BLOB-1931	BUMB (Brown UMass Boston Corpus)
1990s	ICE-GB	ICE-USA

be made about the genres. However, with the much shorter ICE (and with other million-word corpora, such as Brown and LOB, as well), it is an open question whether the forty 2,000-word samples of academic prose contained in the ICE, for instance, are enough to adequately represent this genre. And given the range of variation that Biber (1988) documents in academic prose, forty samples are probably not enough. Does this mean, then, that million-word corpora are not useful for genre studies?

The answer is no: While these corpora are too short for some studies, for frequently occurring grammatical constructions they are quite adequate for making generalizations about a genre. For instance, Meyer (2014) analyzed pseudo-titles and related appositives in the press genre of the four corpora listed in Table 2.5.

Pseudo-titles are constructions such as *rock vocalist Iggy Pop*, which are similar to equivalent appositives (*a rock vocalist, Iggy Pop*), except that there is no comma pause between the first unit in the appositive and the second unit. While pseudo-titles are very common in American newspapers, they are stigmatized in British newspapers, occurring more frequently in tabloids than in broadsheets.

Meyer (2014) had two goals in his study: to trace the development of appositives in press reportage in general, and to document the rise of pseudo-titles. To conduct this study, he examined four corpora, each containing twenty 2,000-word samples, resulting in a corpus that was 160,000 words in length. ICE-GB and ICE-USA contained newspapers published in the 1990s; BLOB-1931 (which is modeled after the LOB corpus), and BUMB (the Brown UMass Boston Corpus) contained collections of texts taken from American and British newspapers published in the 1930s.

All told, Meyer (2014: 241) found 843 examples of the constructions under study, providing more than enough data to reach statistically significant differences upon which to base generalizations. For instance, Meyer (2014: 246) was able to document a statistically significant shift in the placement of the unit containing the proper noun from the first unit in newspapers from the 1930s:

Teddy Nash, a former Swindon Town goalkeeper, shone against his old club, some of his saves, especially early in the match, being cleverly accomplished. (BLOB A07)

to the second unit in newspapers published in the 1990s:

Zuidmulder was overwhelmed by the Louvre's famous art treasures by ***the Dutch masters, Michelangelo and the Leonardo da Vinci***. (ICE-USA W2C-006)

This shift, in turn, led to the rise of pseudo-titles, constructions in which the determiner is deleted in the first part of the pseudo-title (Meyer 2014: 246):

Economist John Kendall at Baring Brothers said: "The people who get this right are those who can assess the information when war breaks out." (ICE-GB W2C-013)

Studying the amount of internal linguistic variation in a genre is one way to determine how many samples of a genre need to be included in a corpus; applying the principles of sampling methodology is another way.

Social scientists have developed a sophisticated methodology based on mathematical principles that enables a researcher to determine how many "elements" from a "sampling frame" need to be selected to produce a "representative" and therefore "valid" sample. A sampling frame is determined by identifying a specific population that one wishes to make generalizations about. For instance, in planning the creation of the Santa Barbara Corpus of Spoken American English, it was decided that recordings of spontaneous conversations would include a wide range of speakers from around the United States representing, for instance, different regions of the country, ethnic groups, and genders. In addition, speakers would be recorded using language in a variety of different contexts, such as conversing over the phone, lecturing in a classroom, giving a sermon, and telling a story (www.linguistics.ucsb.edu/research/santa-barbara-corpus). The idea behind this sampling frame was to ensure that the corpus contained samples of speech representing the ways that people speak in different regions of the country and in different contexts.

Social scientists have developed mathematical formulas that enable a researcher to calculate the number of samples they will need to take from a sampling frame to produce a representative sample of the frame. Kretzschmar, Meyer, and Ingegneri (1997) used one of

Kalton's (1983: 82) formulas to calculate the number of books published in 1992 that would have to be sampled to create a representative sample. *Bowker's Publisher's Weekly* lists 49,276 books as having been published in the United States in 1992. Depending on the level of confidence desired, samples from 2,168 to 2,289 books would have to be included in the corpus to produce a representative sample. If each sample from each book was 2,000 words in length, the corpus of books would be between 4,336,000–4,578,000 words in length. Kalton's (1983) formula could be applied to any sampling frame that a corpus compiler identifies, provided that the size of the frame can be precisely determined.

Sampling methodology can also be used to select the particular individuals whose speech and writing will be included in a corpus. For instance, in planning the collection of demographically sampled speech for the BNC, "random location sampling procedures" were used to select individuals whose speech would be recorded (Crowdy 1993: 259). Great Britain was divided into 12 regions. Within these regions, "thirty sampling points" were selected: locations at which recordings would be made and that were selected based on "their ACORN profile. . .(A Classification of Regional Neighbourhoods)" (Crowdy 1993: 260). This profile provides demographic information about the types of people likely to live in certain regions of Great Britain, and the profile helped creators of the BNC to select speakers of various social classes to be recorded. In selecting potential speakers, creators of the BNC also controlled for other variables, such as age and gender.

In using sampling methodology to select texts and speakers and writers for inclusion in a corpus, a researcher can employ two general types of sampling: probability sampling and nonprobability sampling (Kalton 1983). In probability sampling (employed above in selecting speakers for the BNC), the researcher very carefully pre-selects the population to be sampled, using statistical formulas and other demographic information to ensure that the number and type of people being surveyed are truly representative. In non-probability sampling, on the other hand, this careful pre-selection process is not employed. For instance, one can select the population to be surveyed through the process of "*haphazard*, *convenience*, or *accidental sampling*" (Kalton 1983: 90); that is, one samples individuals who happen to be available. Alternatively, one can employ "*judgment*, or *purposive*, or *expert choice*" sampling (Kalton 1983: 91); that is, one decides beforehand who would be best qualified to be sampled (e.g. native

rather than non-native speakers of a language, educated vs. non-educated speakers of a language, etc.). Finally, one can employ "*quota sampling*" (Kalton 1983: 91), and sample certain percentages of certain populations. For instance, once could create a corpus of American English by including in it samples reflecting actual percentages of ethnic groups residing in the United States (e.g. 13.2 percent African Americans, 17 percent Hispanic and Latino Americans, etc.).

Although probability sampling is the most reliable type of sampling, leading to the least amount of bias, for those who created first generation in corpora, this kind of sampling presented considerable logistical challenges. The mathematical formulas used in probability sampling often produce very large sample sizes, as the example above with books illustrated. And there were simply not enough resources available to create corpora beyond the million words in length. However, for many second-generation corpora, such as COCA, GloWbE, or many of the corpora accessible through Sketch Engine (www.sketchengine.co.uk/), size is not an issue, since it currently requires fewer resources to create corpora that contain millions, even billions of words. But for smaller corpora, particularly those containing copyrighted material fully accessible to the user, size becomes a more significant issue, since obtaining copyright permission is a time-consuming and difficult task.

Judgment, or purposive, or expert choice sampling, a second type of sampling, was used to create the Brown Corpus. That is, prior to the creation of the Brown Corpus, it was decided that the writing to be included in the corpus would be randomly selected from collections of edited writing at four locations:

(1) for newspapers, the microfilm files of the New York Public Library;
(2) for detective and romantic fiction, the holdings of the Providence Athenaeum, a private library;
(3) for various ephemeral and popular periodical material, the stock of a large secondhand magazine store in New York City;
(4) for everything else, the holdings of the Brown University Library. (quoted from Francis 1979: 195)

Other corpora have employed other non-probability sampling techniques. The American component of ICE used a combination of "judgment" and "convenience" sampling: Every effort was made to collect speech and writing from a balanced group of constituencies (e.g. equal numbers of males and females), but ultimately what was finally included in the corpus was a consequence of whose speech or

writing could be most easily obtained. For instance, much fiction is published in the form of novels or short stories. However, many publishers require royalty payments from those seeking reproduction rights, a charge that can sometimes involve hundreds of dollars. Consequently, most of the fiction in the American component of ICE consists of unpublished samples of fiction taken from the Internet: fiction for which permission can usually be obtained for no cost and which is available in computerized form. The Santa Barbara Corpus of Spoken American English, as noted earlier, employed a variation of "quota" sampling, making every effort to collect samples of speech from a representative sampling of men and women, and various ethnicities and regions of the United States.

2.10 The Time Frame for Selecting Texts

Most corpora contain samples of speech or writing that have been written or recorded within a specific time frame. Synchronic corpora (i.e. corpora containing samples of English as it is presently spoken and written) contain texts created within a relatively narrow time frame.

For instance, the Brown and LOB Corpora contain written texts published in 1961. The written and spoken texts included within the BNC were published/recorded in the late twentieth century (www.natcorp.ox.ac.uk/). The COCA contains texts created between the years 1990 and 2019, with new texts being added on a yearly basis. The Collins Corpus, used as the basis for creating the COBUILD dictionaries, is a monitor corpus that is currently 4.5 billion words in length (https://collins.co.uk/page/The+Collins+Corpus).

In creating a synchronic corpus, the corpus compiler wants to be sure that the time frame is narrow enough to provide an accurate view of contemporary English undisturbed by language change. However, linguists disagree about whether purely synchronic studies are even possible: New words, for instance, come into the language every day, indicating that language change is a constant process. Moreover, even grammatical constructions can change subtly in a rather short period of time. Leech, Mair, Hundt, and Smith (2009) conducted a corpus study of language change involving such structures as modal verbs and passive constructions that took place in four corpora spanning the years 1961–92 (p. xx):

Brown (1961)
LOB (1961)
F-LOB (1991)
Frown (1992)

The Frown Corpus is modeled after the Brown Corpus, except that it contains texts published in 1992. The F-LOB Corpus is a similar updating of the LOB Corpus. In their analyses of these corpora, they found many changes taking place during the 31-year period that the corpora cover. For instance, they documented changes over this time span in the use of *wh*-relative clauses and *that*-relative clauses. *That*-relative clauses became more common than *wh*-relative clauses (e.g. *the child that was late* vs. *the child who was late*), particularly in American English. They attributed this change to the process of colloquialization: the perception of *wh*-relatives as being restricted in usage to "formal written text types" (p. 230).

With diachronic corpora (i.e. corpora used to study historical periods of English), the time frame for texts is somewhat easier to determine, since the various historical periods of English are fairly well-defined. However, complications can still arise. For instance, Rissanen (1992: 189) notes that one goal of a diachronic corpus is "to give the student an opportunity to map and compare variant fields or variant paradigms in successive diachronic stages in the past." To enable this kind of study in the Helsinki Corpus, Rissanen (1992) remarks that the Old and Early Middle English sections of the corpus were divided into 100-year sub-periods; the Late Middle and Early Modern English periods were divided into 70- to 80-year periods. The length of sub-periods was shorter in the later periods of the corpus "to include the crucial decades of the gradual formation of the 15th-century Chancery standard within one and the same period" (Rissanen 1992: 189). The process that was used to determine the time frames included in the Helsinki Corpus indicates that it is important in diachronic corpora not to just cover pre-determined historical periods of English but to think through how significant events occurring during those periods can be best covered in the particular diachronic corpus being created.

2.11 The Linguistic Background of Speakers and Writers Whose English Is Included in a Corpus

The previous sections provided descriptions of various corpora of English, and the many methodological issues that one must

address both in the creation and analysis of a corpus. But one issue that has not been discussed so far concerns the linguistic backgrounds of those whose speech or writing has been included in a corpus. If one is creating a corpus of, say, spoken American English, should the speech of only native speakers of American English be included? And if the answer is yes, how does one determine exactly what a native speaker is?

As Davies (2003: 16–18) demonstrates, the notion of a native speaker is a much-vexed idea, with labels such as mother tongue, first language, dominant language, and home language being more complex than they appear to be. Many factors have been proposed as indicators of native speaker status, such as the age of acquisition (the notion of a "critical period" early in life), lack of a foreign accent, and overall fluency in speech. But these factors are complicated too. For instance, consider the notion of fluency, as many people who speak English as a second or foreign language speak it quite fluently. In 2011, the United States Census Bureau circulated a questionnaire called the "American Community Survey" (www2.census.gov/library/publications/2013/acs/acs-22/acs-22.pdf). The first question asked respondents whether they spoke a language other than English at home. If they did, they were asked to respond to a series of additional questions. One question asked them to rank the quality of English that they spoke. A total of 60,577,020 respondents replied, with the responses listed as follows:

very well: 58.2 percent
well: 19.4 percent
not well: 15.4 percent
not at all: 7 percent

Of course, self-reports can be problematic, since how one perceives his or her speech can be quite different from how others perceive it. Still, excluding someone's speech in, say, a corpus of American English based on whether they speak English as an additional language has its problems, since arguably, American English can be regarded as the sum total of everyone who speaks it.

It is therefore quite important that those creating corpora explicitly define the population whose speech will be sampled. For instance, all ICE corpora contain texts representing "the English of adults (age 18 or over) who have been educated through the medium of English to at least the end of secondary schooling" (www.ucl.ac.uk/english-usage/projects/ice.htm). Note that this description does not

necessarily exclude bilingual speakers from the corpus. It simply assures that individuals whose speech will be included in the corpus have had exposure to English over a significant period. The BNC is defined as "a monolingual British English corpus: it comprises text samples which are substantially the product of speakers of British English" (www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#spodes). This restriction allows for a certain degree of flexibility as well, as it would permit a variety of different speakers and writers of British English to be represented in the corpus.

In non-native varieties of English, the level of fluency among speakers will vary considerably: as Schmied (1996: 187) observes, "it can be difficult to determine where an interlanguage ends and educated English starts." Nevertheless, in selecting speakers for inclusion in a corpus of non-native speech or writing, it is important to specify specific criteria for selection, such as how many years an individual has used English, in what contexts they have used it, how much education in English they have had, and so forth.

To determine whether an individual's speech or writing is appropriate for inclusion in a corpus (and also to elicit the sociolinguistic variables), one can have individuals contributing texts to a corpus fill out a biographical form in which they supply the information necessary for determining whether their native or non-native speaker status meets the criteria for inclusion in the corpus. For instance, individuals can be asked what languages they speak, how long they have spoken them, and in what contexts they have used them, such as in the home, workplace, school, and so forth. If the residence history on the biographical form is unclear, it is also possible to interview the individual afterwards, provided that he or she can be located; if the individual does not fit the criteria for inclusion, his or her text can be discarded.

Determining native or non-native speaker status from authors of published writing can be considerably more difficult, since it is often not possible to locate authors and have them fill out biographical forms. In addition, it can be misleading to use an individual's name alone to determine native speaker status, since someone with a non-English sounding name may have immigrant parents and nevertheless be a native speaker of English, and many individuals with English sounding names may not be native speakers: One of the written samples of the American component of ICE had to be discarded when the author of one of the articles called on the telephone and explained that he was not a native speaker of American English but Australian English. Schmied

(1996: 187) discovered that a major newspaper in Kenya had a chief editor and training editor who were both Irish and who exerted considerable editorial influence over the final form of articles written by native Kenyan reporters. This discovery led Schmied (1996) to conclude that there are real questions about the authenticity of the English used in African newspapers. Ultimately, however, when dealing with written texts, one simply must acknowledge that in compiling a corpus of American English, for instance, there is a chance that the corpus will contain some non-native speakers, despite one's best efforts to collect only native-speaker speech.

In spoken dialogues, one may find out that one or more of the speakers in a conversation do not meet the criteria for inclusion because they are not a native speaker of the variety being collected. However, this does not necessarily mean that the text must be excluded from the corpus, since there is annotation that can be included in a corpus indicating that certain sections of a sample are "extra-corpus" material; that is, material not considered part of the corpus for purposes of word counts, generating KWIC (key word in context), and so forth.

2.12 Controlling for Sociolinguistic Variables

There are a variety of sociolinguistic variables that will need to be considered before selecting the speakers and writers whose texts are being considered for inclusion in a corpus. Some of these variables apply to the collection of both spoken and written texts; others are more particular to spoken texts. In general, when selecting individuals whose texts will be included in a corpus, it is important to consider the implications that their gender, age, and level of education will have on the ultimate composition of the corpus. For the spoken parts of a corpus, several additional variables need to be considered: the dialects the individuals speak, the contexts in which they speak, and the relationships they have with those they are speaking with. The potential influences that these variables have on a corpus are summarized in the following categories.

2.12.1 Gender Balance

It is relatively easy when collecting speech and writing to keep track of the number of males and females from whom texts are being collected. Information on gender (which is defined biologically

in corpus linguistics) can be requested on a biographical form, and in written texts, one can often tell the gender of an individual by his or her first name.

Achieving gender balance in a corpus involves more than simply ensuring that half the speakers and writers in a corpus are female and half male. In certain written genres, such as scientific writing, it is often difficult to achieve gender balance because writers in these genres are predominantly male – an unfortunate reality of modern society. To attempt to collect an equal proportion of writing from males and females might actually misrepresent the kind of writing found in these genres. Likewise, in earlier periods, men were more likely to be literate than women and thus to produce more writing than women. To introduce more writing by females into a corpus of an earlier period distorts the linguistic reality of the period. A further complication is that much writing, particularly scientific writing, is co-written, and if males and females collaborate, it will be difficult to determine precisely whose writing is actually represented in a sample. One could collect only articles written by a single author, but this again might lead to a misrepresentation of the type of writing typically found in a genre. Finally, even though an article may be written by a female or a male, there is no way of determining how much an editor has intervened in the writing of an article and thus distorted the effect that the gender of the author has had on the language used in the article.

In speech, other complications concerning gender arise. Research has shown that gender plays a crucial role in language usage. For instance, women will speak differently with other women than they will with men. Consequently, to adequately reflect gender differences in language usage, it is best to include in a corpus a variety of different types of conversations involving males and females: women speaking only with other women, men speaking only with other men, two women speaking with a single man, two women and two men speaking, and so forth.

To summarize, there is no one way to deal with all the variables affecting the gender balance of a corpus. The best that the corpus compiler can do is to be aware of the variables, confront them head on, and deal with them as much as is possible during the construction of a corpus.

2.12.2 Age

There are ranges of age groups that have been included in the many corpora that have been created. For instance, there are

special-purpose corpora containing the speech of individuals up to the age of 16. The Child Language Data Exchange System, or CHILDES Corpus, includes transcriptions of children engaging in spontaneous conversations in English and other languages. The Bergen Corpus of London Teenager English (COLT) contains the conversations of adolescents aged 13–17 years. The Polytechnic of Wales Corpus contains transcriptions of conversations between children (aged 6–12 years) and a "friendly" adult concerning their "favourite games or TV programmes."

Overall, though, corpora have tended to contain the speech of adults, largely because to collect the speech of children and adolescents, one often must obtain the permission not just of the individual being recorded but of his or her parents as well, a complicating factor in an already complicated endeavor.

2.12.3 Dialect Variation

It is also important to consider the extent to which a corpus should contain a range of dialects, both social and regional, that exist in any language.

In many respects, those creating historical corpora have been more successful in representing regional variation than those creating modern-day corpora: The regional dialect boundaries in Old and Middle English are fairly well-established, and in the written documents of these periods, variant spellings reflecting differences in pronunciation can be used to posit regional dialect boundaries. For instance, Rissanen (1992: 190–2) is able to describe regional variation in the distribution of *(n)aught* (meaning "anything," "something," or "nothing") in the Helsinki Corpus because in Old and Early Middle English this word had variant spellings reflecting different pronunciations: a spelling with <a> in West-Saxon (e.g. Old English *(n)awuht*) and a spelling with <o> in Anglian or Mercian (e.g. Old English *(no)whit*). Social variation is more difficult to document because, as Nevalainen (2000: 40) notes, "Early modern people's ability to write was confined to the higher social ranks and professional men." Therefore, it is unavoidable that sociolinguistic information in a historical corpus will either be unavailable or skewed towards a particular social class.

Because writing is now quite standardized, it no longer contains traces of regional pronunciations. However, even though the modern-day corpus linguist has access to individuals speaking many different

regional and social varieties of English, it is a significant undertaking to create a spoken corpus that is balanced by region and social class. If one considers only American English, a number of different regional dialects can be identified, and within these major dialect regions, one can isolate numerous sub-dialects (e.g. Boston English within the coastal New England dialect). If social dialects are added to the mix of regional dialects, even more variation can be found, as a social dialect such as African-American Vernacular English can be found in all major urban areas of the United States. In short, there are numerous dialects in the United States, and to attempt to include representative samplings of each of these dialects in the spoken part of a corpus is nothing short of a methodological nightmare.

What does one do, then, to ensure that the spoken part of a corpus contains a balance of different dialects? In selecting speakers for inclusion in the BNC, twelve dialect regions were identified in Great Britain, and from these dialect regions, 100 adults of varying social classes were randomly selected as those whose speech would be included in the corpus (Crowdy 1993: 259–60). Unfortunately, the speech of only 100 individuals can hardly be expected to represent the diversity of social and regional variation in a country the size of Great Britain. Part of the failure of modern-day corpora to adequately represent regional and social variation is that creators of these corpora have had unrealistic expectations. As Chafe, Du Bois, and Thompson (1991: 69) note, the thought of a corpus of "American English" to some individuals conjures up images of "a body of data that would document the full sweep of the language, encompassing dialectal diversity across regions, social classes, ethnic groups. . .[enabling] massive correlations of social attributes with linguistic features." But to enable studies of this magnitude, the corpus creator would have to have access to resources far beyond those that are currently available – resources that would enable the speech of thousands of individuals to be recorded and then transcribed.

Because it is not logistically feasible in large countries such as the United States or Great Britain to create corpora that are balanced by region and social class, some corpus linguists have devoted their energies to the creation of corpora that focus on smaller dialect regions. Tagliamonte (1998) and Tagliamonte and Lawrence (2000: 325–6), for instance, contain linguistic discussions based on a 1.5 million-word corpus of York English that has been subjected to extensive analysis and that has yielded valuable information on dialect patterns (both social and regional) particular to this region of

England. Kirk (1992) has created the Northern Ireland Transcribed Corpus of Speech, a corpus containing transcriptions of interviews of speakers of Hiberno English.

2.12.4 Social Contexts and Social Relationships

Speech takes place in many different social contexts and among speakers between whom many different social relationships exist. When we work, for instance, our conversations take place in a specific and very common social context – the workplace – and among speakers of varying types: equals (e.g. co-workers), between whom a balance of power exists, and disparates (e.g. an employer and an employee), between whom an imbalance of power exists. Because the employer has more power, he or she is considered a "superordinate" in contrast to the employee, who would be considered a "subordinate." At home (another social context), other social relationships exist: a mother and her child are not simply disparates but intimates as well.

There is a vast amount of research that has documented how the structure of speech is influenced by both the social context in which speech occurs and the social relationships existing between speakers. As Biber and Burges (2000) note, to study the influence of gender on speech, one needs to consider not just the gender of the individual speaking but the gender of the individual(s) to whom a person is speaking. Because spontaneous dialogues will constitute a large section of any corpus containing speech, it is important to make provisions for collecting these dialogues in as many different social contexts as possible. The London-Lund Corpus contains an extensive collection of spoken texts representing various social relationships between speakers: spontaneous conversations or discussions between equals and between disparates; radio discussions and conversations between equals; interviews and conversations between disparates; and telephone conversations between equals and between disparates (Greenbaum and Svartvik 1990: 20–40).

The various components of the ICE (www.ice-corpora.uzh.ch/en/design.html) contain spontaneous conversations taking place in many different social contexts, ranging from face-to-face conversations to classroom discussions. The Michigan Corpus of Academic Spoken English (MICASE) contains samples of academic speech occurring in many different academic contexts, such as lectures given by professors to students as well as conversations between students in study

groups. This ensured that the ultimate corpus created would represent the broad range of speech contexts in which academic speech occurs (Simpson, Lucka, and Ovens 2000: 48). In short, the more diversity one adds to the social contexts in which language takes place, the more assured one can be that the full range of contexts will be covered.

2.13 Mega Corpora

The previous sections have discussed several methodological issues that need to be considered as one plans and creates a corpus. But as corpora have grown larger, it has become a much more complicated undertaking to ensure that a corpus is both balanced and representative. This is especially the case with web-based corpora, which are quite large in size and whose content is sometimes difficult to determine. For instance, at 1.9 billion words in length, the Corpus of Global Web-Based English is so lengthy that it would be impossible to determine not just the content of the corpus but the distribution of such variables as the gender of contributors, their ages, and so forth.

At the other end of the spectrum are those who would question the necessity of highly planned corpora such as the BNC. Kilgarriff, Atkins, and Rundell (2007) argue that such factors as the easy availability of online texts such as newspapers, as well as services for transcribing recorded speech, have made the creation of corpora much larger than the BNC less arduous. But perhaps a middle ground is a more reasonable approach to take: the creation of so-called small and beautiful corpora for those who need a balanced and representative corpus, and large mega corpora for those whose research needs require extensive amounts of data.

2.14 Conclusions

To create a valid and representative corpus, it is important, as this chapter has shown, to carefully plan the construction of a corpus before the collection of data even begins. This process is guided by the ultimate use of the corpus. If one is planning to create a multi-purpose corpus, for instance, it will be important to consider the types of genres to be included in the corpus; the length not just of

the corpus but of the samples to be included in it; the proportion of speech versus writing that will be included; the educational level, gender, and dialect backgrounds of speakers and writers included in the corpus; and the types of contexts from which samples will be taken. However, because it is virtually impossible for the creators of corpora to anticipate what their corpora will ultimately be used for, it is also the responsibility of the corpus user to make sure that the corpus he or she plans to conduct a linguistic analysis of is a valid corpus for the particular analysis being conducted. This shared responsibility will ensure that corpora become the most effective tools possible for linguistic research.

3 Building and Annotating a Corpus

Once the basic outlines of a corpus are determined, it is time to begin the actual creation of the corpus. This is a two-part process, involving the collection and encoding of data.

Collecting data involves:

(1) recording speech (e.g. spontaneous conversations, classroom discussions), or obtaining second-hand recordings/transcriptions of, for instance, speeches, news broadcasts, and other types of spoken language that are widely available;
(2) gathering written texts, either in digital or print formats;
(3) obtaining permission to use texts from speakers, writers, or copyright holders;
(4) keeping careful records about the texts collected and the individuals from whom they were obtained.

Encoding data involves:

(1) transcribing speech;
(2) scanning written texts if digitized versions are not available;
(3) adding (as resources permit) three different types of information into the corpus (McEnery and Hardie 2012: 29–31):

Metadata: This is purely descriptive information about a given corpus. For instance, in a spoken corpus, one could include separately from the corpus itself such information as the size of the corpus and the types of speech contained in it (e.g. dialogues or monologues) as well as demographic information about the individual speakers in the corpus, such as their age, ethnicity, gender, and so forth.

Textual markup: This type of markup provides various kinds of information about the internal structure of a corpus. For instance, in a spoken corpus consisting of spontaneous dialogues, markup can be inserted into the corpus to identify which conversant is speaking; to mark segments of overlapping speech (i.e. places in a conversation where individuals speak simultaneously); to note where speakers pause and the length of their pauses; and so forth. In a written text, markup can be inserted to demarcate paragraph boundaries, to identify words that are

in italics or boldface, and to mark other features that are particular to written texts.

Linguistic annotation: The process of annotating a corpus involves running software that can (1) *tag* a corpus (add part-of-speech tags to all of the words in the corpus, such as nouns, prepositions, and verbs), or (2) *parse* a corpus (add markup that identifies larger structures, such as verb phrases, prepositional phrases, and adverbials). Grammatical markup is inserted when a corpus is tagged or parsed.

Metadata is a key component of any corpus: users need to know precisely what is in a corpus. *Textual markup* is important too, though corpora will vary in terms of how much of such markup they contain. For instance, marking segments of overlapping speech can be a time-consuming process. Consequently, some spoken corpora may not mark where speakers overlap. *Linguistic annotation* varies from corpus to corpus as well. While it is quite common for corpora to be lexically tagged, parsing a corpus is a much more complicated process. Therefore, relatively few corpora have been parsed.

Even though the processes of collecting and encoding data for inclusion in a corpus are described above as separate processes, in many senses they are closely connected: After a conversation is recorded, for instance, it may prove more efficient to transcribe it immediately, since whoever made the recording will be available either to transcribe the conversation or to answer questions about the recording to aid in its transcription. If a text is collected, and saved for later computerization and annotation, the individual who collected the text may not be around to answer questions, and information about the text may consequently be lost. Of course, logistical constraints may necessitate collecting texts at one stage and computerizing and annotating them at a later stage, in which case it is crucial that as much information about the text be obtained initially so that, at a later stage, those working with the text will be able to easily recover this information. The kinds of information to collect will be discussed in greater detail in Section 3.2.

3.1 General Considerations

As discussed in Chapter 2, before the actual data for a corpus is collected, it is important to carefully plan exactly what will be included in the corpus: the kinds and amounts of speech and/or writing, for instance, as well the range of individuals whose speech

and writing will become part of the corpus. Once these determinations are made, the corpus compiler can begin to collect the actual speech and writing to be included in the corpus. However, it is important not to become too rigidly invested in the initial corpus design, since obstacles and complications may be encountered while collecting data that may require changes in the initial corpus design: It might not be possible, for instance, to obtain recordings for all the genres originally planned for inclusion in the corpus, or copyright restrictions might make it difficult to obtain certain kinds of writing. In these cases, changes are natural and inevitable and, if they are carefully made, the integrity of the corpus will not be compromised.

The International Corpus of English (ICE) Project provides a number of examples of logistical realities that forced changes in the initial corpus design for some of the components. After the project began, it was discovered that not all the regional groups involved in the project would be able to collect all of the text categories originally planned for inclusion in the corpus. For instance, in ICE-East Africa (which includes texts collected in Kenya and Tanzania), it was not possible to collect examples of scripted monologues, such as legal presentations and scripted commentaries. However, to make up for this deficiency, additional texts were collected in the other categories: 120 scripted monologues (e.g. news broadcasts and speeches) instead of the 50 samples required in the other components of the ICE (http://clu.uni.no/icame/manuals/ICE_EA.PDF). Adaptations such as this ensured that each component would contain one million words of speech and writing.

3.2 Collecting Samples of Speech

Speech is the primary mode of human communication. As a result, there are various types of speech: not just spontaneous multiparty dialogues but scripted and unscripted monologues, radio and television interviews, telephone conversations, class lectures, and so forth. Given the wealth of speech that exists, as well as the logistical difficulties involved in recording and transcribing it, collecting data for the spoken part of a corpus is much more labor-intensive than collecting written samples. As Adolphs and Carter (2013: 8) note, "Spoken corpus development is very expensive in both time and real costs." Drawing upon work that they did creating the CANCODE Corpus, they comment that a one-hour recording of casual

conversation will yield around 10,000 words of text (p. 9). And if the amount of time it takes to transcribe and annotate this one hour of speech is factored in, creating a full corpus of spontaneous conversations is a very labor-intensive undertaking.

In collecting any kind of speech, the central concern is obtaining speech that is "natural." This is a particularly important issue when gathering spontaneous multi-party dialogues, such as informal talks between two or more individuals. If multi-party dialogues are not carefully collected, the result can be a series of recordings containing very stilted and unnatural speech. Collecting "natural" multi-party dialogues involves more than simply recording people as they converse. As anyone who has collected language data knows, if speakers know that their speech is being recorded, they will change the way that they speak, a phenomenon called the "observer's paradox" (Labov 1972): when individuals know that their speech is being monitored, they may adjust their speaking style and attempt to speak what they perceive to be "correct" English – a style of speaking that they would not use in a normal informal conversation.

To avoid this problem, those creating earlier corpora, such as the London-Lund Corpus, recorded people surreptitiously, and only after recordings were secretly made were individuals informed that they had been recorded. While it may have been acceptable and legal back in the1950s and 1960s to record individuals without their knowledge, now such recordings are not only considered unethical within the scientific community but may in fact be illegal in many locales. It is therefore imperative both to inform individuals that their speech is being recorded and to obtain written permission from them to use their speech in a corpus. This can be accomplished by having individuals being recorded sign a release form prior to being recorded. In addition, many universities and other organizations have Institutional Review Boards that review any research conducted on "human subjects" by faculty members and will grant approval for use of such subjects only if accepted practices are followed.

Since it is not possible to include surreptitious speech in a corpus, does this mean that non-surreptitiously gathered speech is not natural? This is an open question, since it is not possible to answer it with any definite certainty. However, there are ways to increase the probability that that the speech included in a corpus will be natural and realistic.

First, before individuals are recorded, they should be given in written form a brief description of the project in which they are

participating. In this description, the purpose of the project should be described, and it should be stressed that speech is being collected for descriptive linguistic research, not to determine whether those being recorded are speaking "correct" or "incorrect" English. In a sense, these individuals need to be given a brief introduction to a central tenet of modern linguistics: that no instance of speech is linguistically better or worse than any other instance of speech, and that all types of speech are legitimate, whether they are perceived as standard or non-standard.

A second way to enhance naturalness is to record as lengthy a conversation as possible so that when the conversation is transcribed, the transcriber can select the most natural segment of speech from a much lengthier speech sample, for instance, 30 minutes or longer. This length of conversation increases the probability that a natural and coherent sample of speech can be extracted from this longer sample. The initial part of the conversation can then be discarded, since people are sometimes nervous and hesitant upon first being recorded but after a while become less self-conscious and start speaking more naturally. Moreover, a lengthier segment allows the corpus compiler to select a more coherent and unified segment of speech to ultimately include in the corpus.

How much speech needs to be recorded is also determined by the type of speech being recorded. Spontaneous dialogues, for instance, may require lengthier segments of recorded speech because of features such as hesitations, pauses, and interruptions – all of which slow down the pace of speech. On the other hand, monologues (especially scripted monologues) contain far fewer pauses and hesitations. Thus, less speech is needed to reach these necessary numbers of words for the particular corpus being created.

When it comes time to actually make recordings, whoever is making the recordings needs to follow a few basic principles of recording to ensure the most natural recordings as possible. Probably the least desirable way to make a recording is to have the research assistant sitting silently nearby with microphone in hand while the people being recorded converse. This all but ensures that the individuals being recorded will constantly be reminded that they are part of a "linguistic experiment." As much as possible, those making recordings should try to record individuals in the actual environments in which they typically speak, such as the family dinner table, the workplace, restaurants, the car, informal get-togethers and so forth.

Because the goal is to record people in natural speaking environments, it is best for the research assistant not to be present during recordings. He or she can simply set up the recording equipment, turn on the recorder, and leave. Alternatively, the people being recorded can be loaned the recording equipment and taught to set it up and turn on the recorder themselves. This latter option was used to gather speech in the demographic component of the British National Corpus (BNC). Individuals participating in the project were given portable tape recorders and an adequate supply of cassettes, and were instructed to record all of the conversations they had for periods ranging from two to seven days (Crowdy 1993: 260; www.natcorp.ox.ac.uk/docs/URG/BNCdes.html#body.1_div.1_div.5_div.1). To keep track of the conversations they had, participants filled out a logbook, indicating when and where the recordings occurred as well as who was recorded. This method of collection habituates participants to the process of being recorded and, additionally, ensures that a substantial amount of speech is collected.

In certain natural speaking environments, such as restaurants or automobiles, there will often be a considerable amount of ambient noise, resulting in recordings containing variable amounts of inaudible speech that cannot be accurately transcribed. This problem can be handled with annotation during the actual transcription of the recording that marks certain sections as inaudible. It can also sometimes be minimized using commonly available audio editing software. If high-quality recordings are desired, individuals can be recorded in an actual recording studio, as was done in collecting speech samples for the Map Task Corpus, a corpus of conversations between individuals giving directions to various locations (see Thompson, Anderson, and Bader 1995; for more detail, see http://groups.inf.ed.ac.uk/maptask/s). However, in speech collected this way, while high-quality recordings will be obtained, often the naturalness of the speech collected will be compromised. Therefore, it is probably wise to sacrifice recording quality in favor of natural speech.

While collecting natural speech is a key issue when recording multi-party dialogues, it is less of an issue with other types of speech. For instance, those participating in radio and television broadcasts will undoubtedly be conscious of the way they are speaking, and therefore may heavily monitor what they say. However, heavily monitored speech is "natural" speech in this context. Therefore, this is precisely the kind of speech one wants to gather. Other types of spoken language, such as public speeches (especially if they are scripted), are also heavily edited.

When recording any kind of spoken English, it is important to consider the quality of the audio recorder and microphone to be used. In earlier corpora, most corpus compilers used analog recorders and cassette tapes, since such recorders were small and unobtrusive and cassette tapes were inexpensive. However, with the rise of digital technology, there are now a variety of recorders that are widely available, such as digital audiotape (DAT) recorders as well as Minidisc recorders. These recorders make high-quality digital recordings, which can then be exported and used in special software programs that aid in the transcription of speech.

It is equally important to consider the quality and type of microphone to be used to make recordings. A low-quality microphone will produce recordings that are "tinny" even if a good tape recorder is used. It is therefore advisable to invest resources in good microphones, and to obtain microphones that are appropriate for the kinds of recordings being made. To record a single individual, it is quite acceptable to use a traditional uni-directional microphone, a microphone that records an individual speaking directly into the microphone. For larger groups, however, it is better to use omni-directional microphones: microphones that can record individuals sitting at various angles from the microphone. Lavaliere microphones, which are worn around the neck, are useful for recording individuals who might be moving around when they speak, as people lecturing to classes or giving speeches often do. Wireless microphones are appropriate in recording situations of this type too, and avoid the problem of the speaker being constrained by the length of the cord attaching the microphone and recorder. There are also extra-sensitive microphones for recording individuals who are not close to the microphone, as is the case in a class discussion, where those being recorded are spread out all over a room and might not be close enough to a traditional microphone for an audible recording to be made. For recording telephone conversations, special adapters can be purchased that record directly off landline telephones. For cell phones, there are apps that can be used to record conversations. High-quality microphones can be fairly expensive, but they are worth the investment.

To find information on specific recorders and microphones, it is useful to consult what researchers who specialize in the study of spoken language use to make voice recordings. For instance, The Division of Psychology and Language Sciences at University College London provides a listing of recorders and microphones used by individuals doing research in areas such as phonetics and speech pathology (www

.phon.ucl.ac.uk/resource/audio/recording.html). Anthropologists and ethnographers who do fieldwork also make use of recording equipment for recording spoken language (https://medium.com/@vanhoben/ethnographic-fieldwork-equipment-that-hopefully-wont-break-the-bank-new-digital-tools-6068ddd869f5).

Even with the best recording equipment, however, assembling a large number of recordings suitable for inclusion in a corpus is a time-consuming and sometimes frustrating process: for every 10 recordings made, it may turn out that some of them are unusable. For instance, conversants might be speaking naturally for certain periods of time, only to stop and say something irrelevant, such as "Is the recorder still working?" Remarks such as this illustrate that no matter how hard one tries, it is impossible to make many individuals forget that their speech is being monitored and recorded. Other problems include excessive background noise that makes all or part of a conversation inaudible, or people who are given a recorder to record their speech and then operate it improperly, in some cases not recording any of the conversation they set out to record. Those compiling spoken corpora should therefore expect to gather much more speech than they will actually use to compensate for all the recordings they make that contain imperfections preventing their use in the ultimate corpus being created.

While most of the recordings for a corpus will be obtained by recording individuals with microphones, many corpora will contain sections of broadcast speech. This type of speech is best recorded not with a microphone but directly from a radio or television by running a cord from the audio output plug on the radio or television to the audio input plug on the tape recorder. It is also possible to get a wide variety of written transcriptions of broadcast speech from such sources as talk shows, interviews, and news conferences. But using such second-hand sources raises questions about the accuracy of the transcriptions: the extent to which what is transcribed accurately matches what was actually said.

Inaccuracies in such transcriptions can potentially be a problem. Molin (2007), for instance, found that in the Hansard Corpus of Canadian parliamentary proceedings, transcriptions of the proceedings did not always capture the precise language that was used by conversants. In many cases, changes were made so that usages that parliamentarians actually uttered conformed more to prescriptive norms: contracted forms were changed to full forms (e.g. *don't* to *do not*) or *be going to* was replaced with *will* (p. 207). In contrast, in a

corpus containing transcripts from the broadcast network CNN, Hoffmann (2007: 71) was confident that the transcripts closely matched the speech that was transcribed. He noted, for instance, that features of discourse, such as *oh* and *well*, commonly occurred in the transcriptions, thus suggesting that the transcripts closely represented the actual utterances of speakers (though certainly a much more detailed analysis of transcripts and recordings would be needed to fully verify this claim).

A perusal of several transcripts and recordings made by the broadcast network NPR (www.npr.org/templates/transcript/transcript.php?storyId=15166387) revealed occasional mistranscriptions, but overall, a close fidelity with the actual language transcribed. One of the few errors observed involved transcribing *react* in the example below instead of the form actually used in the recording: *interact*.

It's also changed the way they **react** with doctors, their families, and even with strangers.

But it is important to note to that there will always be errors in any transcription, whether it is based directly on a recording or taken from a transcription made available by a second party. The ultimate goal is to get as much accuracy as possible and be practical. Therefore, the time saved in using second-party transcripts is worth the tolerance of a certain level of error, provided that some checking is done to ensure overall accuracy in the transcripts included in a corpus. The process of transcribing speech will be discussed in greater detail in 3.x.

3.3 Collecting Samples of Writing

Although collecting samples of writing is considerably less complicated than collecting samples of speech, one significant obstacle is encountered when collecting writing: copyright restrictions. Under the "fair use" provisions of current U.S. copyright laws, it is possible in certain circumstances to use copyrighted material without receiving the explicit permission from the copyright holder. However, the "circumstances" are stated rather generally and are subject to interpretation (www.copyright.gov/fair-use/more-info.html). Thus, including something in a corpus without getting explicit permission from the copyright holder could involve copyright infringement.

In many first-generation corpora, such as the Brown and BNC, clearance was obtained for all copyrighted material. But because of

the huge size of recent mega-corpora, obtaining such clearance is simply not possible. With the Corpus of Contemporary American English (COCA), though, workarounds were implemented that allowed users full access to the corpus without violating copyright law. For instance, searches of the corpus retrieve only "very short 'Keyword in Context' displays, where users see just a handful of words to the left and the right of the word(s) searched for." Additionally, the corpus can be used only for academic research (www.english-corpora.org/copyright.asp).

If attempts are made to secure permission to use copyrighted material in a corpus, it is best to over collect the number of texts to include in each part of the corpus: A written text may be gathered for possible inclusion in a corpus, and its author (or publisher) may not give permission for its use, or (as was a common experience gathering written texts for inclusion in the American component of ICE) contact is made requesting permission but no reply is ever received.

Because of the difficulties in obtaining permission to use copyrighted materials, most corpus compilers have found themselves collecting far more written material than they are able to obtain permission to use: for ICE-USA, permission had been obtained to use only about 25 percent of the written texts initially considered for inclusion in the corpus. Moreover, some authors and publishers will ask for money to use their material: one publisher requested half an American dollar per word for a 2,000-word sample of fiction considered for inclusion in ICE-USA! Therefore, if a corpus is to be used only for non-profit academic research, this should be clearly stated in any letter of inquiry or email requesting permission to use copyrighted material and many publishers will sometimes waive fees. However, if there will be any commercial use of the corpus, special arrangements will have to be made both with publishers supplying copyrighted texts and those making use of the corpus.

In gathering written texts for inclusion in a corpus, the corpus compiler will undoubtedly have a predetermined number of texts to collect within a range of given genres: twenty 2,000-word samples of fiction, for instance, or ten 2,000-word samples of learned humanistic writing. However, because there is so much writing available, it is sometimes difficult to determine precisely where to begin to locate texts. Since most corpora are restricted to a certain time frame, this frame will of course narrow the range of texts, but even within this time frame, there is an entire universe of writing.

In earlier corpora, written texts needed to be scanned or re-typed to be converted into digital formats. Currently, however, most written texts are also available in digital formats. For instance, newspapers and magazines are accessible in digital formats that can be easily used in corpora. Many commercial publications, such as novels, are also available digitally, but often in formats that cannot be copied.

3.4 Keeping Records of Texts Gathered

As written texts are collected and spoken texts are recorded, it is imperative that accurate records are kept about the texts and the writers and speakers that created them. For ICE-USA, research assistants filled out a written checklist supplying specific information for each spoken and written text that was collected.

First of all, each text was assigned a number that designated a specific genre in the corpus in which the sample would be potentially included. For instance, a text numbered S1A-001 would be the first sample considered for inclusion in the genre of "direct conversations"; a text numbered S1B-001, on the other hand, would be the first sample considered for inclusion in the genre of "classroom lectures." A numbering system of this type (described in detail in Greenbaum 1996b: 601–14) allows the corpus compiler to keep easy record of where a text belongs in a corpus and how many samples have been collected for that part of the corpus. After a text was numbered, it was given a short name providing descriptive information about the sample. In ICE-Great Britain, sample S1A-001 (a spontaneous dialogue) was named "Instructor and dance student, Middlesex Polytechnic" and sample S1B-001 (a class lecture) was entitled "Jewish and Hebrew Studies, 3rd year, UCL." The names supplied to a text sample are short and mnemonic and give the corpus compiler (and future users of the corpus) an idea of the type of text that the sample contains.

The remaining information recorded about texts depended very much on the type of text that was being collected. For each spoken text, a record was kept of when the text was recorded, where the recording took place, who was recorded, who did the recording, and how long the recording was. For each person recorded, a short excerpt of something they said near the start of the recording was written down so that whoever transcribed the conversation would be able to match the speech being transcribed with the speaker. It can be

very difficult for a transcriber to do this if he or she has only the recording to work with and must figure out who is speaking. For speech samples recorded from television or radio, additional information was written down, such as what station the recording was made from, where the station is located, and who should be contacted to obtain written permission to use the sample. For written texts, a complete bibliographical citation for the text was recorded along with the address of the publisher or editorial office from which permission to use the written text could be obtained.

In addition to keeping records of the texts that are recorded, it is equally important to obtain ethnographic information from individuals contributing either a sample of speech or a written text. The particular information collected will very much depend on the kind of corpus being created and the variables that future users of the corpus will want to investigate. Because ICE-USA is a general-purpose corpus, only fairly general ethnographic information was obtained from contributors: their age, gender, occupation, and a listing of the places where they had lived over the course of their lives. Other corpora have kept different information on individuals, relevant to the particular corpus being created. The Michigan Corpus of Academic Spoken English (MICASE) collected samples of spoken language in an academic context. Therefore, not just the age and gender of speakers in the corpus were recorded but their academic discipline (e.g. humanities and arts, biological and health sciences), academic level (e.g. junior undergraduates, senior faculty), native-speaker status, and first language (Simpson and Swales 2001: 35). More recent mega corpora, such as COCA, tend to contain no ethnographic information about speakers, largely because the large size of the corpora makes tracking down ethnographic information about speakers and writers virtually impossible.

Ethnographic information is important because those using the ultimate corpus that is created might wish to investigate whether gender, for instance, affects conversational style, or whether younger individuals speak differently than older individuals. It is important for these researchers to be able to associate variables such as these with specific instances of speech. While it is relatively easy to obtain ethnographic information from individuals being recorded (they can simply fill out the form when they sign the permission form), tracking down writers and speakers on radio or television shows can be very difficult. Therefore, it is unrealistic to expect that ethnographic information will be available for every writer and speaker in a corpus. And

indeed, many current corpora, such as the BNC and ICE, contain missing information on many speakers and writers.

After all the above information is collected, it can be very useful to enter it into a database, which will allow the progress of the corpus to be tracked. This database can contain not just information taken from the forms described above but other information as well, such as whether the text has been computerized yet, whether it has been proofread, and so forth. Creating a corpus is a huge undertaking, and after texts are collected, it is very easy to file them away and forget about them. It is therefore crucial to the success of any corpus undertaking that accurate information be kept about each text to be considered for inclusion in the corpus.

3.5 Encoding Spoken and Written Data

Encoding spoken data is a much more complicated process than encoding written data.

To encode spoken data, recordings of speech first need to be transcribed, an extremely lengthy process requiring the transcriber to listen to the same segments of speech repeatedly until an accurate transcription is achieved. Although the process of transcription has been automated, current voice recognition technology has not reached the level of sophistication to be able to accurately transcribe the most common type of speech: spontaneous conversations.

In earlier corpora, written texts, which existed in printed form, had to be optically scanned into a computer – a process that often produced digitized texts with numerous scanning errors that had to be manually corrected. However, so many written texts are now available in digitized formats that scanning is mainly restricted to texts dating back to the pre-electronic era.

There are a number of general considerations to bear in mind when beginning the process of computerizing both spoken and written texts. First, because texts to be included in a corpus will be edited with some kind of text editing program, it may be tempting to save computerized texts in a file format used by a word processing program (such as files with the extension .doc in Microsoft Word). However, these files will be incompatible with any of the programs customarily used with corpora, such as concordancing programs.

In earlier corpora, the standard was the ASCII (or text) file format, a format that had both advantages and disadvantages. The main

advantage at the time was that ASCII was a universally recognized text format that could be used with any word processing program and the numerous software programs designed to work on corpora, such as taggers and parsers and concordances. The disadvantage was that because ASCII has a fairly limited set of characters, many characters and symbols cannot be represented in it. The creators of the Helsinki Corpus had to develop a series of special symbols to represent characters in earlier periods of English that are not part of the ASCII character set: the Old English word "ðæt" ("that"), for instance, is encoded in the corpus as "+t+at" with the symbol "+t" corresponding to the Old English thorn "ð" and the symbol "+a" to the Old English ash "**æ**" (Kytö 1996). This system of symbols is specific to the Helsinki Corpus. The successor to ASCII, Unicode, eliminates the need for the creation of ad hoc symbols because it contains an expanded character-set that is able to represent virtually all of the characters found in the languages of the world.

When creating a corpus, it is easiest to save individual texts in separate files stored in directories that reflect the hierarchical structure of the corpus. This does not commit one to distributing a corpus in this format: the ICAME CD-ROM (2nd ed.) allows users to work with an entire corpus saved in a single file. But organizing a corpus into a series of directories and sub-directories makes working with the corpus much easier, and allows the corpus compiler to keep track of the progress being made on corpus as it is being created. Figure 3.1 contains a partial directory structure for regional corpora included in the ICE.

Each ICE component consists of texts in two main directories – one containing all the spoken texts included in the corpus, the other all the written texts. These two directories, in turn, are divided into a series of sub-directories containing the main types of speech and writing that were collected: the spoken part into monologues and dialogues, the written part into printed and non-printed material.

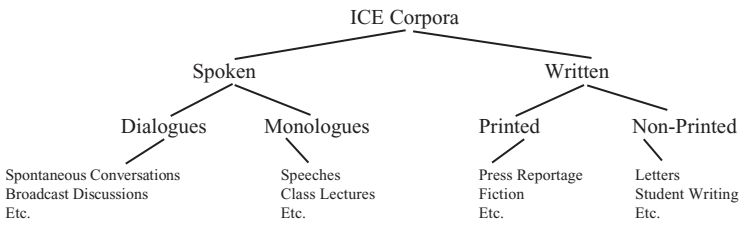

Figure 3.1 Sample directory structure

Each text included in a given ICE component is assigned an identification letter and number indicating the type of speech or writing that it represented. For instance, for the 90 texts labeled S1A-001 to S1A-090, the letter S indicates that each 2,000-word sample represents spoken English; the numerals 001-090 that it was a private conversation (either a spontaneous conversation or a telephone call); and the uppercase A that it was a dialogue. While this numbering system is unique to ICE components, a similar system can be developed for any corpus project.

Other directories can be created to fit the needs of the research team building a particular corpus. For instance, a "draft" directory is useful for early stages of corpus development and can contain spoken texts that are in the process of being transcribed or written texts that have been computerized but not proofread. Once a draft version of a text contains "metadata" and "textual markup", it can then be placed into a "lexical (pending proofreading)" directory to indicate that the text will be ready for use as a lexical version of the corpus once it has been proofread. There are then two stages of proofreading. The first stage involves proofreading each individual sample in a corpus with samples being placed into a "proofread 1" directory. A second round of proofreading is done after completion of the entire corpus so that the corpus can be viewed as a whole.

While a text is being worked on at a particular stage of analysis, it receives an additional file extension to indicate that work on the text is in progress. For instance, while a draft version of a text in the category of business transactions is being created, the text is saved as "S1B-071di", the "i" indicating that work on the text is incomplete. As a particular text is being worked on, a log is maintained that notes what work was done on the text and what work needs to be done. At each stage of analysis, to avoid duplication of work, it is most efficient to have a single person work on a text; at the proofreading stage, it is best to have the text proofread by someone not involved with any prior version of the text, since he or she will bring a "fresh" perspective to the text.

Finally, although inserting "structural" markup into a text is separate from the process of actually computerizing the text, there are many instances where markup can be inserted while texts are being computerized. For instance, in transcribing spontaneous conversations, the transcriber will encounter numerous instances of overlapping speech – individuals speaking at the same time. The segments of speech that overlap need to be marked so that the eventual user of the

corpus knows which parts overlap in the event that he or she wishes to study overlapping speech. If annotating overlapping segments of speech is done separately from the actual transcription of the text, the individual doing the annotation will have to go through the tape repeatedly to reconstruct the overlaps – a process that could be done more efficiently by the person doing the transcription. Likewise, speaker identification tags – tags indicating who is speaking – are more efficiently inserted during the transcription of texts. With written texts, if two line breaks are inserted between paragraphs while a text is being computerized, then paragraph tags can be inserted automatically at a later stage. Of course, some markup is probably better inserted after a text sample is computerized. But because computerizing and annotating a text is such an integrated process, it is best to combine the processes whenever this is possible.

But there are caveats to the process of creating a corpus outlined in this section. As corpora have become larger and larger, oftentimes containing millions of words of text, the feasibility of, for instance, proofreading a corpus or placing individual samples into neatly delineated sections becomes less viable: Such corpora are simply too large for any kind of proofreading to be done, or for a single text type (such as a press editorial from a particular newspaper) to be placed into a single directory.

3.6 Transcribing Speech

Traditionally, speech had been transcribed using a special transcription machine that has a foot pedal that stops and starts a cassette tape and also automatically rewinds the tape to replay a previous segment. As anyone who has ever transcribed speech knows, the flow of speech is much faster than the ability of the transcriber to type. Therefore, it is extremely important to have the capability of automatically replaying segments.

Because of recent advances in computer technology, it is now possible to use software programs designed specifically to transcribe samples of speech that have been digitized. "VoiceWalker 3.0b" was developed to aid in the transcription of texts included within the Santa Barbara Corpus of Spoken American English. "SoundScriber" is a similar program used to transcribe texts that are part of the MICASE (www-personal.umich.edu/~ebreck/code/sscriber/). Both of these programs are available at the above URLs as freeware,

and work very much like cassette transcription machines: the transcriber opens a word processing program in one window and the transcription program in another. After a sample of digitized speech is loaded into the program, short segments of the sample can be automatically replayed until an accurate transcription is achieved. These programs accept as input a variety of different audio formats, including WAV files as well as MP3 files. And for files in different audio formats, the open source program Audacity can be used to convert many different audio files into WAV or MP3 files (www.audacityteam.org/). This program can also be used to convert cassette tapes to digital formats.

While transcription machines and software can ease the process of transcribing speech, there is no getting around the fact that speech must be manually transcribed. And while there are no immediate prospects that this process will be automated, speech recognition programs have improved considerably in recent years, and, in the near future, offer the hope that they can at least partially automate the process of transcribing speech.

Early transcription programs, such as Dragon Dictate (https://en.wikipedia.org/wiki/DragonDictate), offered little hope to transcribers, since they required words in a recording to be carefully pronounced and followed by a pause. With a little training, such programs produced reasonable speech recognition but only of a very artificial type of speech. In the early 1990s, a major technological breakthrough occurred: the ability of speech recognition programs to transcribe continuous speech. There now exist a number of programs, such as Dragon NaturallySpeaking (https://en.wikipedia.org/wiki/Dragon_NaturallySpeaking), that have large lexicons and with training can quite accurately recognize carefully articulated monologues.

More recently, there have been advances in the development of programs that can automatically transcribe samples of digitized speech. However, the accuracy of such transcriptions depends upon the type of speech that is being transcribed. Monologues (particularly those that are scripted rather than spontaneous) are the easiest types of speech to automatically transcribe because the voice of a single person is all that needs to be recognized. In contrast, spontaneous unscripted dialogues are much more difficult to transcribe because such speech contains multiple speakers whose conversations contain, for instance, hesitations, overlaps, re-formulations, and incomplete sentences.

But despite these difficulties, there has been progress in developing programs that may in the future help automate the transcriptions of

spontaneous dialogues. Xiong et al. (2017) describe a voice recognition system that they developed that was able to quite accurately transcribe spoken conversations in two corpora: the Switchboard Corpus contains 2,400 phone conversations (totaling 260 hours of speech) taking place between two individuals, and the CallHome Corpus consists of 120 thirty-minute phone conversations between intimates: either friends or members of the same family.

To test their system, Xiong et al. (2017) first had the conversations in both corpora transcribed by professional transcribers. They found that the transcription error rate for the Switchboard Corpus was 5.9 percent and for the CallHome Corpus 11.3 percent. They then had the automated system that they developed transcribe the same conversations. The error rates were slightly lower: 5.8 percent for the Switchboard Corpus and 11.0 percent for the CallHome Corpus. Although this system works with only two speakers, it does foreshadow a time when the transcription of speech for inclusion in linguistic corpora can be greatly expedited.

Transcribing speech is in essence a highly artificial process, since an exclusively oral form of language is represented in written form. Consequently, before any transcription is undertaken, it is important to decide just how much that exists in a spoken text one wishes to include in a transcription of it. Compilers of corpora have varied considerably in how much detail they have included in their transcriptions of speech. The spoken sections of the COCA do not contain speech that was recorded and transcribed in-house. Instead, spoken samples in this corpus consist entirely of transcriptions of numerous radio and television programs that were created by a third-party. Given the cost and effort of creating a corpus of speech, it is understandable why corpora of this type exist, and while they do not contain spontaneous face-to-face conversations, they do provide a substantial amount of spoken language occurring in other speech-based registers.

At the other extreme are corpora of speech that attempt to replicate in a transcription as much information as possible about the particular text being transcribed. The Santa Barbara Corpus of Spoken American English, for instance, contains not only an exact transcription of the text of a spoken conversation (including hesitations, repetitions, partially uttered words, and so forth) but also annotation marking various features of intonation in the text, such as tone unit boundaries, pauses, and pitch contours. This kind of detail is included because creators of this corpus attached a high value to the

importance of intonation in speech. The main drawback of this kind of detailed transcription is the lengthy amount of time it takes to annotate a text with information about intonation. The advantage is that very detailed information about a spoken text is provided to the user, thus ensuring that a broad range of studies can be conducted on the corpus without any doubt about the authenticity of the data.

Whether one does a minimal or detailed transcription of speech, it is important to realize that it is not possible to record all the subtleties of speech in a written transcription. As Cook (1995: 37) notes, a spoken text is made meaningful by more than the words one finds in a transcription: How a conversation is interpreted depends crucially upon such contextual features as paralanguage (e.g. gestures and facial expressions), the knowledge the participants have about the cultural context in which the conversation takes place, their attitudes towards one another, and so forth. All of this extra-linguistic information is very difficult to encode in a written transcription without the corpus compiler developing an elaborate system of markup identifying this information and the transcriber spending hours both interpreting what is going on in a conversation and inserting the relevant markup. It is therefore advisable when transcribing speech to find a middle ground: to provide an accurate transcription of what people actually said in a conversation, and then, if resources permit, to add extra information (e.g. marking for various features of intonation).

In reaching this middle ground, it is useful to follow Chafe's (1995) principles governing the transcription of speech. A transcription system, Chafe (1995: 55) argues, should (1) adopt as much as possible standard conventions of orthography and "capitalize on habits [of literacy] already formed" by corpus users; (2) strive to be as iconic as possible; and (3) be compatible with current computer technology. The following sections contain a brief survey of some of the issues that the corpus creator must address in creating a satisfactory system for transcribing speech. The survey is not exhaustive but illustrative, since sketching out a complete system of transcription is beyond the scope of the present discussion.

3.7 A General Overview of Metadata, Textual Markup, and Linguistic Annotation

As noted earlier, in order to make a corpus "usable" for linguistic analysis, it is necessary for the corpus compiler to insert

Metadata, Textual Markup, and Linguistic Annotation into the corpus. Metadata and Textual Markup are more *descriptive*, noting, respectively, what is in a corpus and where, for instance, paragraph boundaries in a written corpus occur or speakers pause in a spoken corpus. Linguistic Annotation is more *grammatical* in nature, providing linguistic information about the various structures occurring within a particular corpus. For instance, if a corpus is lexically tagged, each word in the corpus is assigned a part-of-speech designation, such as noun, verb, preposition, and so forth. In contrast, if a corpus is syntactically parsed, various types of grammatical information is provided, such as which structures are noun phrases, verb phrases, subordinate clauses, and imperative sentences.

The following sections describe the three different ways of describing the content of a corpus, beginning with a brief overview of Metadata and more detailed descriptions of Textual Markup and Linguistic Annotation.

3.8 Metadata and Textual Markup

In earlier corpora, there was no standardized system of indicating in a spoken corpus, for instance, sequences of overlapping speech: places in a transcription where the speech of two or more individuals overlaps. Likewise, because written corpora were encoded in text files, there was no way to indicate certain features of orthography, such as italicization or boldface fonts. Consequently, different corpora contained different kinds of markup to describe the same linguistic phenomena. More recently, however, the Text Encoding Initiative (TEI) has developed a standardized system of annotation not just for linguistic corpora (www.tei-c.org/release/doc/tei-p5-doc/en/html/CC.html) but for electronic texts in general (www.tei-c.org/). Metadata and textual markup are important because they can narrow the range of possible choices when a corpus is searched.

One important feature of a TEI-conformant document is what is called the TEI header (www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html), which supplies Metadata about a particular electronic document. Although such headers can get quite complicated, one important part of the header is the file description, which provides information about a document, as illustrated below in a header for The Written Component of ICE-USA:

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>The Written Component of ICE-USA</title>
    <respStmt>
      <resp>Prepared by</resp>
      <name>Charles F. Meyer</name>
      <name>Hongyin Tao</name>
    </respStmt>
    <publicationStmt>
  <distributor>University of Zurich</distributor>
  <address>
    <addrLine>http://www.ice-corpora.uzh.ch/en.
      html</addrLine>
  </address>
    <availability>
    <p>Freely available on a non-commercial basis.</p>
    </availability>
    </publicationStmt>
  <projectDesc>
    <p>400,000 words of various types of written
      American English published in the 1990s</p>
  </projectDesc>
  </fileDesc>
</teiHeader>
```

(www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-fileDesc.html)

Each part of the header is nested. For instance, the first part of the header, **<teiHeader>**, is enclosed in angle brackets; the last part of the header, **</teiHeader>**, is also enclosed in angle brackets but with a / (slash) after the first bracket to indicate that all the information between the first header and second header are part of the header. Other such relationships can be found throughout the header: the title of the document, *The Written Component of ICE-USA*, is enclosed with an open marker, **<title>**, and a close marker, **</title>**. Various levels of indentation are also used to illustrate the hierarchy.

There is also TEI-conformant Textual Markup to describe features occurring within a particular corpus. For instance, an unscripted conversation will contain features of speech such as speaker turns, pauses, or partially articulated words. In contrast, a written text will contain features of orthography, such as paragraph boundaries or font changes. The next section will describe this annotation as it applies to

spoken language; a later section will focus on the annotation used in written texts.

3.9 Textual Markup for Features of Spoken Texts

While the particular words of a conversation are easy to transcribe using standard orthography, there are other features of speech (e.g. overlapping speech) for which particular Textual Markup is necessary. Because of the complexity of the TEI system of annotation for speech, the discussion will be more illustrative than comprehensive. This is in line with Hardie's (2014) notion that the average corpus linguist does not need to master the vast complexity of the TEI system of annotation but would be better served to focus instead on essential markup necessary for describing the basic features of a given corpus. In addition, alternatives to the TEI system for representing speech will also be described.

3.9.1 Utterances

Unlike written texts, spoken texts do not always contain grammatically well-formed sentences. For instance, in the conversational excerpt below (taken from the East African component of ICE), the first speaker (B) leaves out the subject, I, before *think*, and does not finish the sentence, instead pausing with *uh* before the next speaker begins talking. In the first part of the second turn, the speaker (C) utters only a partial sentence, the negative particle *Not* followed by a prepositional phrase: *out of uh the question*:

<B> <u>think Mr Juma wants to say something maybe uh</u>
<C> <u> Not out of uh the question </u> <u>What you're trying to discuss now is about the current situation and the near future of about the position of the</u>
<B> <u> the political situation</u>
<C> <u>Yeah and the position of Zanzibar President then is that what you're discussing</u>

All the units are marked with the tag <u>, which indicates that each construction is an utterance: a sequence of words that has meaning, even though it is not a grammatically well-formed sentence. In written texts, which overwhelmingly contain grammatical sentences, a different set of tags would be used: <s> and </s>, which mark the beginning and end of a sentence. Note too that each of the speaker turns receives a

speaker identification tag: a capital letter (B or C) within angle brackets < >. In spoken corpora, the only punctuation marks typically used are apostrophes for contractions and possessives, capitalization of proper nouns, and hyphens for hyphenated words.

3.9.2 Vocalized Pauses and Other Lexicalized Expressions

Speech contains a group of one- or two-syllable utterances that are communicative but not fully lexical in nature. In the TEI system, these utterances are characterized as non-lexical (e.g. snorts, giggles, laughs, coughs, sneezes, or burps) or semi lexical (e.g. *ah, aha, er, huh, ooh, oops, uh, uh-huh, uh-uh, um*) (www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html#TSBAVO). Because there is no universally agreed upon spelling for these expressions, the TEI system provides ways to mark their occurrence in a conversation. For instance, the example below marks a section of a conversation in which a speaker coughs, a non-lexical expression:

```
<vocal>
        <desc> coughs </desc>
</vocal>
```

The next example contains the semi-lexical expression *uh*, which is very common in speech and serves the function of allowing the speaker to think of something that he or she wishes to say next in a conversation:

```
<vocal>
        <desc> uh </desc>
</vocal>
```

It is also common in speech to find examples of expressions that are spelled as two separate words, but pronounced as one word. For instance, *got to*, *have to*, and *going to* are commonly pronounced as, respectively, *gotta*, *hafta*, and *gonna*. Additional examples include *kinda, sorta*, and *lotsa*, which are shortened forms of *kind of*, *sort of*, and *lots of*, respectively. Practice will vary, but typically in corpora where the distinction is made, the form that matches the pronunciation is the one that will be transcribed.

3.9.3 Partially Uttered Words and Repetitions

Speech (especially unscripted speech) contains several false starts and hesitations resulting in words that are sometimes not

completely uttered. In the example below, the speaker begins uttering the preposition *in* but only pronounces the vowel beginning the word.

<$D> There are more this year than <.> i </.> in in your year weren't there (ICE-GB)

In the ICE Project, such incomplete utterances are given an orthographic spelling that best reflects the pronunciation of the incompletely uttered word, and then the incomplete utterance is enclosed in markup, <.> i </.>, that labels the expression as an instance of an incomplete word.

Repetitions can be handled in a similar manner. When speaking, an individual will often repeat a word more than once as he or she is planning what to say next. In the example below, the speaker repeats the noun phrase *the police boat* twice before she completes the utterance:

<$B> <}_><-_>the police boat<-/> <=_>the police boat<=/> <}/_>we know but the warden's boat we don't you know he could just be a in a little rowboat fishing and (ICE-USA)

To accurately transcribe the above utterance, the transcriber will want to include both instances of the noun phrase. However, this will have the unfortunate consequence of skewing a lexical analysis of the corpus in which this utterance occurs, since all instances of *the, police*, and *boat* will be counted twice. To prevent this, the ICE Project has special markup that encloses the entire sequence of repetitions (<}_><}/_>) and then places special markup (<=_>the police boat<=/>) around the last instance of the repetition, the only instance counted in analyses done by ICECUP, the text analysis program used in the ICE Project.

3.9.4 Unintelligible Speech

Very often when people speak, their speech is unintelligible. If two people speak simultaneously, for instance, they may drown out each other's words and the speech of both speakers will become unintelligible. Anyone doing a transcription of spoken dialogues will therefore encounter instances where speech cannot be transcribed because it is not understandable. In the example below (taken from ICE-USA), the TEI tags <unclear> and </unclear > surround the word *them* because the transcriber was uncertain whether this was actually the word the speaker uttered.

<u> What was Moses doing going off in <unclear> them </unclear> jeans </u>

3.9.5 Changing the Names of Individuals Referred to in Spoken Texts

In any conversation, speakers will address themselves by name, and they will talk about third-party individuals, sometimes in unflattering ways – one spoken sample from the American component of ICE contains two brothers talking quite disparagingly about their parents.

In transcribing a recording taken from a public broadcast, such as a radio talk show, it is of little concern whether the actual names of individuals are included in a transcription, since such a conversation was intended for public distribution. In transcribing private conversations between individuals, however, it is crucial that names be changed in transcriptions to protect the privacy of the individuals conversing and the people they are conversing about. Moreover, many universities and other organizations have strict rules about the use of human subjects in research and the extent to which their anonymity must be preserved. And if the recordings accompanying the transcriptions are to be made available as well, any references to people's names (except in publicly available recordings) will have to be edited out of the recordings – something that can be done quite easily with software that can be used to edit digitized samples of speech. In changing names in transcriptions, one can simply arbitrarily substitute new names appropriate to the gender of the individual being referred to or, as was done in the London-Lund Corpus, substitute "fictitious" names that are "prosodically equivalent to the originals" (Greenbaum and Svartvik 1990: 19).

3.9.6 Iconicity and Speech Transcription

Because writing is linear, it is not difficult to preserve the "look" of a printed text that is converted into an electronic document and transferred from computer to computer in text format: although font changes are lost, markup can be inserted to mark these changes; double-spaces can be inserted to separate paragraphs; and standard punctuation (e.g. periods and commas) can be preserved. However, as one listens to the flow of a conversation, it becomes quite obvious that speech is not linear: Speakers very often talk simultaneously, and

while someone is taking their turn in a conversation, another party may fill in brief gaps in the turn with backchannels, expressions such as *yea* and *right* that tell the speaker that his or her speech is being actively listened to and supported. Attempting to transcribe speech of this nature in a purely linear manner is not only difficult but potentially misleading to future users of the corpus, especially if they have access only to the transcription of the conversation, and not the recording.

To explore the many options that have evolved for making transcriptions more iconic, it is instructive, first of all, to view how conversations were transcribed by early conversational analysts, whose transcriptions occurred mainly in printed articles, not in computerized corpora, and how using this early convention of transcription in computerized corpora has certain disadvantages. The conversational excerpt below contains a system of transcription typical of early systems.

A: it's figure three that we have to edit now

[ ]

B: no it's figure
Four we already did figure three

[ ]

A: oh yeah I remember we did it before

(Blachman, Meyer, and Morris 1996, 61).

In the excerpt above, the brackets placed above and below segments of speech in successive speaker turns indicate overlapping segments of speech. For instance, speaker A's uttering of "to edit now" overlaps with speaker B's uttering of "no it's figure." However, instead of representing these overlaps with markup (as a TEI-conformant system would do (cf. www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html#TSSAPA), this system attempts to indicate them iconically by vertically aligning the parts of the conversation that overlap to give the reader of the conversation a sense of how the flow of the conversation took place. While such visual representations of speech are appealing, the implementation of such a system, as Blachman, Meyer, and Morris (1996) note, is very problematical. To vertically align segments of speech, it is necessary to insert tabs into a text. However, as a text gets transferred from computer to computer, font sizes can change and throw the vertical alignment off, giving the corpus user an erroneous representation of the segments of speech that overlap.

To overcome problems like this Blachman, Meyer, and Morris (1996) advocate that segments of a conversation containing overlapping speech be represented in tabular form, a manner of presentation that provides both an accurate and iconic representation of a conversation. Below is how the above conversational excerpt would be represented in a system of this type.

A	B
it's figure three that we have to edit now	
	no it's fig ure four
oh yeah I remember we did it before	we already did figure three

Blachman, Meyer, and Morris (1996: 62)

In the excerpt above, segments of speech in adjoining cells of the table overlap: *edit now* in speaker A's turn, for instance, overlaps with *no it's fig* in speaker B's turn.

An alternative way to represent iconically not just overlapping speech but the flow of conversation in general is to lay it out as though it were a musical score. In the HIAT system (Ehlich 1993), a speaker's contribution to a conversation is represented on a single horizontal line. When the speaker is not conversing, his or her line contains blank space. If speakers overlap, their overlaps occur when they occupy the same horizontal space. In the example below, T begins speaking and, midway through his utterance of the word *then*, H overlaps her speech with T's. For Speakers S1, S2, and Sy, lines are blank because they are not contributing to the conversation at this stage.

T: at once, then the same res/ No, leave it! Would've been (immediately) the same result.
(instantly)
H: Shall I (wipe it out)?
S1:
S2:
Sy:

Ehlich (1993: 134)

Other attempts at iconicity in the above excerpt include the use of the slash in T's turn to indicate that H's overlap is an interruption.

Slashes, according to Ehlich (1993: 128), help convey a sense of "jerkiness" in a speech. Periods and commas mark various lengths of pauses, and the parentheses indicate uncertain transcriptions.

While iconicity is a worthy goal to strive for in transcriptions of speech, it is not essential. As long as transcriptions contain clearly identified speakers and speaker turns, and appropriate annotation to mark the various features of speech, a transcription will be perfectly usable. Moreover, many users of corpora containing speech will be interested not in how the speech is laid out but in automatically extracting information from it. Biber's (1988) study of speech and writing, for instance, was based on tagged versions of the London-Lund and LOB corpora, and through the implementation of a series of algorithms, Biber was able to extract much valuable information from these corpora, without having to examine them manually and make sense out of them with all markup that they contained.

3.10 Computerizing Written Texts

Because written texts are primarily linear in structure, they can easily be encoded in text format: Most features of standard orthography, such as punctuation, can be maintained, and those features requiring some kind of description can be annotated with a TEI-conformant tag. For instance, in the example below, the tag "hi" indicates that the word *very* is highlighted in this context because it is italicized (www.tei-c.org/release/doc/tei-p5-doc/en/html/examples-emph.html).

The child was <hi rend="italics">very</hi> tired.

But while many features of writing can be annotated in a written corpus with TEI-conformant markup, annotating all of these features may prove to be unnecessary. While many users of a spoken corpus will potentially be interested in analyzing overlapping speech, there will likely be very little interest in studying italicized words, for instance, in a written corpus. Consequently, much of what could potentially be annotated in a written text is likely to be of marginal interest to future users of the corpus.

Earlier computer corpora, such as the Brown Corpus, contained texts taken from printed sources, such as newspapers, magazines, and books. To computerize texts from these sources, the texts had to be either keyed in by hand or optically scanned. Nowadays, however, so

many written texts exist in digital formats that they can be easily adapted for inclusion in a corpus. For instance, the 14 billion-word iWeb Corpus consists entirely of texts taken from websites.

However, if one is creating a historical corpus and thus working with texts from earlier periods of English, converting a text into an electronic format can be a formidable task and, in addition, raise methodological concerns that the corpus linguist working with modern texts does not need to consider.

Because written texts from earlier periods may exist only in manuscript form, they cannot be optically scanned but must be typed in manually. Moreover, manuscripts can be illegible in sections, requiring the corpus creator to reconstruct what the writer might have written. Describing a manuscript extract of the Middle English religious work *Hali Meidenhad*, Marcus (1997: 211–12) details numerous complications that the corpus creator would encounter in attempting to create a computerized version of this manuscript: spelling inconsistencies (e.g. <v> and <u> are used interchangeably), accent marks over certain vowels, and colons and dots marking prosodic groupings rather than syntactic constructions.

Although only two versions of *Hali Meidenhad* have survived, thus reducing the level of difference between various versions of this text that the corpus compiler would have to consider, other texts, such as *Ancrene Riwle*, can be found in numerous manuscript editions:11 versions in English, 4 in Latin, and 2 in French (Marcus 1997: 212). Because the corpus compiler is concerned with absolute fidelity to the original, having more than one version of a single text raises obvious methodological concerns that can be dealt with in a number of different ways.

In theory, the corpus compiler could create a corpus containing every manuscript version of a text that exists, and then either let the user decide which version(s) to analyze, or provide some kind of interface allowing the user to compare the various versions of a given manuscript. The Canterbury Project gives users access to all 80 versions of Chaucer's *Canterbury Tales* and allows various kinds of comparisons between the differing versions (Robinson 1998; see also: www.dhi.ac.uk/projects/canterbury-tales/). Because of the enormous amount of work involved in computerizing all versions of a particular text, it is much more practical to computerize only one version. One possibility is to computerize an edited version of a manuscript, that is, a version created by someone who has gone through the various manuscript versions, made decisions concerning

how variations in the differing manuscript versions ought to be reconciled, and produced, in a sense, a version that never existed. A variation on this second alternative (and one advocated by Markus 1997) is for the corpus compiler to become editor and normalize the text as he or she computerizes it, making decisions about which variant spellings to consider, which diacritics to include, and so forth. From a modern perspective, none of these alternatives are ideal, but when working with texts from earlier periods, the corpus compiler must make compromises given the kinds of texts that exist.

One of the earlier and more well-established historical corpora of English is the Helsinki Corpus. This corpus contains texts representing three periods in the development of English: Old English (850–1150), Middle English (1150–1500), and Early Modern English (1500–1710). The texts included from these periods represent various dialect regions of England, such as West-Saxon in the Old English period and East Midlands in the Middle English period. For texts taken from later periods, some sociolinguistic information is provided about some of the writers, such as their age and social status. Various registers are also included in the corpus, such as law, philosophy, history, and fiction.

In 2011, a TEI-XML version of the corpus was released. This represented the first attempt to create a historical corpus conforming to the standards of TEI. One reason for creating a TEI-XML version of the corpus was to avoid the fate of many corpora using their own annotation systems because "as new systems emerge, older ones in limited use are gradually forgotten and the data is rendered effectively inaccessible" (https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/). While this reasoning applied only to the Helsinki Corpus, it is a point worth considering by all individuals creating a new corpus.

3.11 Linguistic Annotation

Garside, Leech, and McEnery (1997: 2) provide a two-part definition of annotation, characterizing it "as the practice of adding **interpretive**, **linguistic** information to an electronic corpus of spoken and/or written language data" [emphasis in original]. Annotation is both interpretive and linguistic in the sense that what is identified in a tagset as a noun, for instance, will vary from theory to theory. In a sentence such as *The rich control our political system*, some might

interpret *rich* as a noun because it follows the article *The*, while others would argue that it is an adjective with an implied noun following it (e.g. *people*). Because linguists have different interpretations of grammatical categories, the tagsets that they create will reflect these differences.

Of the two types of Linguistic Annotation, word-class annotation is much more common in the field of corpus linguistics than annotation used to mark larger grammatical structures, such as noun phrases or verb phrases. The primary reason for this difference is that it is much easier to automatically assign word class "tags" to individual words in a corpus (nouns, verbs, etc.) than to accurately parse larger and more complex grammatical structures, such as nonfinite adverbial clauses or complex noun phrases containing various types of embedded clauses. Tagging programs are also more widely available than parsing programs and have very high accuracy rates. In their survey of the accuracy of various part of speech taggers, Jatav, Teja, and Bharadwaj (2017; https://arxiv.org/ftp/arxiv/papers/1708/1708.00241.pdf) cite accuracy rates ranging from 72.6 percent to 97.93 percent.

While the focus in this section will be primarily on the most common types of annotation – word class and grammatical annotation – other types of annotation are possible too. Lu (2014: 5) comments that the types of constructions that can be annotated are quite "wide-ranging," and can include structures that are "lexical" (morphology) to the "discoursal" (pragmatics). Gries and Berez (2017) describe a range of different types of annotation for the various kinds of linguistic structures found in corpora, including not just word classes (parts of speech), but syntactic structures (parse trees); semantic categories (such as tense and aspect, semantic roles, word senses); and phonetic and prosodic information.

3.11.1 Word Class Annotation

Over the years, a number of different tagging programs have been developed to insert a variety of different tagsets. The first tagging program was designed in the early 1970s by Greene and Rubin (1971) to assign part-of-speech labels to the Brown Corpus. Out of this program arose the various versions of the CLAWS program, developed at the University of Lancaster initially to tag the LOB Corpus (Leech, Garside, and Atwell 1983) and subsequently to tag the BNC (Garside, Leech, and Sampson 1987; Garside and Smith 1997) as well as a recently updated version of the corpus: BNC2014. Additional

programs for tagging corpora include the Brill Tagger (https://cl.lingfil.uu.se/~bea/publ/megyesi-BrillsPoSTagger.pdf) as well as the Stanford Loglinear Part-of-Speech Tagger (https://nlp.stanford.edu/projects/stat-tagging.shtml).

All of these programs are designed to assign various lexical tags to every word in a corpus. For instance, listed below is a lexically tagged sentence from the BNC2014, the second version of the BNC:

<s n=1000> <w **PNP**>I <w **VVB**>mean <w **DTQ**>what <w **PRP**>about <w **AV0**>apparently <w **PNP**>we <w **VVB**>eat <w **DT0**>more <w NN1>**chocolate** <w CJS>**than** <w DT0>**any** <w AJ0>**other** <w NN1>**country**<c PUN>.
(http://ucrel.lancs.ac.uk/bnc2/bnc2guide.htm)

The tags in this example (enclosed in brackets) are part of what is called the C5 tagset, which consists of 60 tags (http://ucrel.lancs.ac.uk/claws5tags.html).

Although the tags may seem rather idiosyncratic, there were several guiding principles that influenced the various abbreviations that were used. Leech (1997: 25–6) recommends that those creating tagsets strive for "Conciseness," "Perspicuity" (making tag labels as readable as possible), and "Analysability" (ensuring that tags can be "decomposable into their logical parts," with some tags, such as "noun," occurring hierarchically above more specific tags, such as "singular" or "present tense"). For instance, two nouns in the excerpt, *country* and *chocolate*, received the tag **NN1**, which is used to identify a singular common noun. In contrast, a plural common noun, such as *countries*, would receive the tag **NN2**. Notice too that the two verbs, *eat* and *mean*, are tagged **VVB**, indicating that these are verbs in the base form. In fact, verbs are such an important word class that 24 of the 60 tags are used to identify types of verbs, with seven tags alone describing the various forms of the verb *be*. The most recent version of the CLAWS tagset, C7, has expanded the tagset to 160+ tags (http://ucrel.lancs.ac.uk/claws7tags.html). This is the result of technological advances made in developing programs for tagging texts.

While the CLAWS tagsets were developed to facilitate the study of the linguistic structure of various kinds of spoken and written texts, other tagsets were created to enable research in the area of natural language processing (NLP), an area of language inquiry that is more focused on the computational aspects of designing taggers (and also parsers) to annotate and study corpora. One of the more well-known tagsets is the Penn

Treebank Tagset, which contains 36 tags (https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html). Like the CLAWS tagsets, this tagset distinguishes between singular and plural nouns but with slightly different tags: **NN** for singular nouns and **NNS** for plural nouns (instead of **NN1** and **NN2**, respectively).

Programs designed to insert word class tags into corpora are of three types: they can be rule-based, stochastic/probabilistic, or a hybrid of the two previous types. In a rule-based tagger, tags are inserted on the basis of rules of grammar written into the tagger. One of the earlier rule-based taggers was the "TAGGIT" program, designed by Greene and Rubin (1971) to tag the Brown Corpus and described in detail in Francis (1979: 198–206). The first step in the tagging process, Francis (1979) notes, is to look up a given word in the program's lexicon, and if the word is found, it is assigned however many tags associated with the word in the lexicon. If after this search, the word is not found, an attempt is made to match the ending of the word with a list of suffixes and the word class tags associated with the suffixes. If this search also fails to find a match, the word is arbitrarily assigned three tags: singular or mass noun, verb (base form), or adjective – three form class designations into which the majority of words in English will fall.

Of the words reaching this stage of analysis, 61 percent will have one tag, and 51 percent of the remaining words will have suffixes associated with one tag. The remaining words will have more than one tag and are thus candidates for disambiguation. Initially, this is done automatically by a series of "context frame rules" that look to the context in which the word occurs. For instance, in the sentence *The ships are sailing*, the word *ships* will have two tags: plural noun and third person singular verb. The context frame rules will note that *ships* occurs following an article, and will therefore remove the verb tag and assign the tag plural noun tag to this word. Although the context frame rules can disambiguate a number of tags, the process is quite complex, as Francis (1979: 202) observes, because many of the words surrounding a word with multiple tags will have multiple tags themselves, making the process of disambiguation quite complicated. Consequently, 23 percent of the remaining tags had to be manually disambiguated: Analysts had to look at each example and decide which tag was most appropriate, a process that itself was subject to error and inconsistency and that led to further post-editing.

While the TAGGIT program had a relatively low accuracy rate, subsequent rule-based taggers have increased overall accuracy rates

considerably. For instance, the rule-based tagger EngCG-2 (cf. Samuelsson and Voutilainen 1997) was designed to overcome some of the problems in early rule-based taggers like TAGGIT. In particular, rules in EngCG-2 have wider application than in TAGGIT and are able to "refer up to sentence boundaries (rather than the local context alone)" (Voutilainen 1999: 18). This capability has greatly improved the accuracy of EngCG-2 over TAGGIT, with a very high accuracy rate of 99.7 percent (https://pdfs.semanticscholar.org/854d/f0c95726bec38e4a62844d159040c745623b.pdf).

While rule-based taggers rely on rules written into the tagger, other taggers are probabilistic/stochastic; that is, they assign tags based on the statistical likelihood that a given tag will occur in a given context. Garside and Smith (1997: 104) give the example of the construction *the run* beginning a sentence. Because *run* starts the sentence and is preceded by the determiner *the*, there is high probability that *run* is a noun rather than a verb. The advantage of stochastic/probabilistic taggers is that they can be trained on corpora and over time develop very high accuracy rates. However, even though stochastic/probabilistic and some rule-based taggers such as EngCG-2 can achieve accuracy rates exceeding 95 percent, the remaining inaccuracies can be more extensive than one might think. Consequently, some researchers have taken to developing hybrid taggers – taggers that are both stochastic/probabilistic and rule-based and that employ the strengths of both of these types of taggers.

The hybrid tagger CLAWS4 was used to tag the BNC2014, the most current version of the written component of the BNC. This tagger employs "a mixture of probabilistic and non-probabilistic techniques" (http://ucrel.lancs.ac.uk/bnc2/bnc2autotag.htm). One of the earlier stages of tagging involves the assignment of initial tags. For instance, a word such as *paint* would be assigned three different tags to reflect its various meanings: NN1 (singular common noun), VVB (the base form of a lexical verb), and VVI (the infinitive form of a verb). The next stage of analysis would draw upon probabilities to narrow the number of possible tags for the word *paint*. For instance, if *paint* occurred in the sentence *I like to paint*, it would obviously be tagged as VVI (the infinitive form of a verb) because it follows the infinitive marker *to*.

The earlier stages of processing with the CLAWS4 tagger yielded an accuracy rate of 97 percent. While this seems like a very high accuracy rate, in a large corpus, the remaining 3 percent of words with multiple tags can constitute a very sizable number of words. To accurately assign a single tag to these words, two rule-based processes were developed, in

the form of what are termed "template rules." For instance, words such as *after*, *before*, and *since*, which can be either prepositions or subordinating conjunctions, were consistently mis-tagged:

1. I will arrive *after* dinner. [preposition]
2. After the movie, we will go out for drinks. [preposition]
3. I will call you *after* I finish the report. [subordinating conjunction]
4. I will contact the editor *after* reviewing the book contract [subordinating conjunction]

To correctly identify each of these instances of *after* as either a preposition or a subordinating conjunction, the following rule was developed (https://ucrel.lancs.ac.uk/bnc2/bnc2autotag.htm):

#AFTER [CJS^PRP] PRP, ([!#FINITE_VB/VVN])16, #PUNC1

Basically, what this rule states is that *after* is more likely to be a preposition than a subordinating conjunction if after 16 words none of the following constructions appear:

a. a finite verb
b. a verb having the form of a past participle
c. a comma

In example (1) below, *after* is clearly a preposition because no verb follows it in the remainder of the sentence. In example (2), *after* is also a preposition because a comma follows *movie*. In contrast, in example (3), *after* is clearly a subordinating conjunction because a verb, *finish*, follows two words after it. Likewise, *after* is a subordinating conjunction in example (4) because a past participle, *reviewing*, occurs directly after it.

(1). I will arrive *after* dinner. [preposition]

(2). After the movie, we will go out for drinks. [preposition]

(3). I will call you *after* I finish the report. [subordinating conjunction]

(4). I will contact the editor *after* reviewing the book contract. [subordinating conjunction]

As the discussion in this section has shown, tagsets vary considerably in the number of part-of-speech tags that they contain. These variations reflect not just differing conceptions of English grammar but the varying uses that the tags are intended to serve.

The Penn Treebank tagset (www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html) contains 36 tags that provide

information about the basic form classes in English: nouns, adjectives, adverbs, verbs, and so forth. For instance, there are four basic tags for the class of adverbs in English:

RB Adverb (*however, usually, naturally, here, good*)
RBR adverb, comparative (*better*)
RBS adverb, superlative (*best*)
WRB wh-adverb (*when, where*)

Although the Penn Treebank can be used purely to conduct linguistic analyses, its main purpose is to advance research in the area of Natural Language Processing (NLP), an area of research that does not always require a more finely graded system of lexical tags.

For those using corpora to conduct more detailed grammatical analyses, larger tagsets are more desirable because they allow the retrieval of a wide range of grammatical constructions. For instance, the ICE tagset is based on the view of grammar articulated in Quirk et al. (1985). Not only is this grammar very comprehensive, taking into consideration constructions with both low and high frequencies in English, but it contains descriptions of spoken as well as written English (www.ice-corpora.uzh.ch/dam/jcr:63a3259d-62ea-4d97-a475-13d303442005/TaggingManual.pdf). Consequently, the ICE tagset is quite detailed, containing 20 base tags (e.g. major word classes, such as noun or adjective) with each tag annotated for various kinds of features. For instance, there are tags for 8 different classes of adverb with various subdivisions for each of these classes. Three of the types of adverbs are in the class of **General Adverbs**:

thin **ADV(ge)**
thinner **ADV(ge,comp)**
thinnest **ADV(ge,sup)**

The second type of adverb is in the class of **Wh-Adverbs**:

when **ADV(rel)**

The remaining six semantic classes contain tags for classifying various other types of adverbs expressing such notions as additive (**add**) *both/neither*; exclusive (**excl**) *only/merely*; intensifier (**inten**) and *very/too*; particularizer (**partic**) *mainly, in particular.*

In contrast, the Penn Treebank tagset is much smaller (36 tags), mainly because this tagset was developed not necessarily to enable detailed linguistic analyses but to advance research in the area of natural language processing (a point described in detail earlier in this

chapter). Thus, this tagset contains, as noted earlier, only four basic tags for adverbs.

While word-class annotation is very well established in corpus linguistics, there are other types of annotation as well. For instance, semantic tagging involves annotating a corpus with markup that specifies various features of meaning. Archer, Wilson, and Rayson (2002; http://ucrel.lancs.ac.uk/usas/usas_guide.pdf) describe a system of semantic tagging that is hierarchically oriented, containing "21 major discourse fields expanding into 232 category labels." For instance, the field of "Money & Commerce in Industry" is annotated as **I**. Various subdivisions of this more general field are indicated numerically. For instance, the field "money generally" is annotated as **I1**, a field containing prototypical words such as *pound, 10P, afford, annuity*, and so forth. The field itself is divided into subdivisions receiving more specific tags. For instance, words within the category "money: affluence" would be tagged **I1.1** and include examples such as *affluence, afford, bank card*, and *cash flow*.

Gries and Berez (2017: 382–90) likewise expand the range of different types of annotation to include semantic annotation, phonetic and phonological annotation, prosodic annotation, sign language and gesture annotation, and interactional intonation.

3.12 Parsing a Corpus

Tagging has become a very common practice in corpus linguistics, largely because taggers have evolved to the point where they are highly accurate: many taggers can automatically tag a corpus (with no human intervention) at accuracy rates exceeding 95 percent. Parsing programs, on the other hand, have variable accuracy rates, largely because it is computationally much more difficult to analyze larger structures, such as phrases and clauses, than individual lexical items. Copestake (2016: 500, note 1) also mentions the difficulty of evaluating the accuracy of parsers. She notes that accuracy rates can reach 90 percent "when trained and tested on newspaper text." But other text types can prove more difficult to parse, resulting in lower accuracy rates.

In corpus linguistics, parsed corpora serve two purposes: to enable the analysis of larger syntactic structures, such as phrases and clauses, by individuals conducting linguistic analyses, and to provide testbeds for those in the area of natural language processing interested in the

development of parsers. This is not necessarily to suggest that these are mutually exclusive categories. For instance, while the Penn Treebank was created by linguists interested in the design of parsers, it can also be used to study the particular parsed linguistic constructions in the 2,499 articles that it contains from the *Wall Street Journal* (https://catalog.ldc.upenn.edu/LDC99T42). Like taggers, parsers can be rule-based, probabilistic, or a hybrid of these two approaches.

One of the most extensively parsed corpora used for linguistic research is ICE-GB, the British component of ICE. Each component of ICE contains a variety of different genres of speech and writing, ranging from spontaneous conversations to scientific writing. Because these genres are so linguistically heterogeneous, a rather complicated methodology was developed to parse ICE-GB.

As was noted in an earlier section, ICE-GB contains a comprehensive set of lexical tags. As Figure 3.1 illustrates, these tags were integrated with tags identifying larger structures, such as phrases and clauses, producing a trichotomy of functions, categories, and features.

Figure 3.2 contains a parse tree for a sentence resulting from a search for the word *lucky* in ICE-GB: *Oh that's very lucky.* The first level of description – functions – is specified both within the clause and the phrase. For instance, *that* is functioning as "subject" (SU); "*s*" as verb (VB, a contracted form of *is*); and *very lucky* as subject complement (CS). Categories are represented at both the phrase and word level: *that* is a "noun phrase" (NP), *rainfall* a "noun" (N) that is also "head" of the noun phrase (NPHD). Features describe various characteristics of functions or categories. For instance, the noun *rainfall* is a "common" noun (com) that is "singular" (sing).

Wallis (2003) describes the process of parsing ICE-GB. First, the corpus was parsed using the TOSCA Parser (Aarts, van Halteren, and

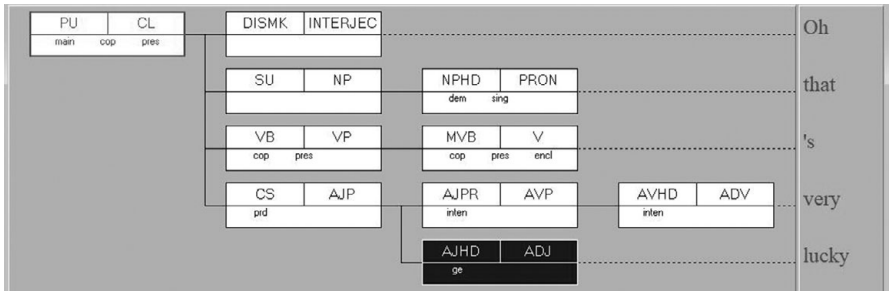

Figure 3.2 Parse tree from ICE-GB

Oostdijk 1996), a rule-based parser. Because this parser "produces a number of alternative parses" (Wallis 2003: 63), the various parses had to be individually inspected to determine which was the correct parse. Moreover, 25 percent of the text units within ICE-GB were not parsed. To parse these units, a second parser, the Survey Parser (Fang, 1996), was used, which produced mixed results as well. Overall, the production of an accurately parsed version of ICE-GB necessitated a considerable amount of manual intervention.

Following the release of ICE-GB, a second parsed corpus using the same architecture was created: the Diachronic Corpus of Present-Day Spoken English (DCPSE) (www.ucl.ac.uk/english-usage/projects/dcpse/index.htm). This corpus contains 400,000 words of spoken English from ICE-GB and an additional 400,000 spoken words from the London-Lund Corpus, a corpus that was based on texts recorded between 1960 and 1980. Because of the complexity of parsing a corpus, there are relatively few corpora parsed in as much detail as ICE-GB and the DCPSE.

Much more common than fully parsed balanced corpora are Treebanks: corpora that are syntactically or semantically parsed but that lack the genre variation typically found in corpora used to conduct linguistic analyses. For instance, one of the more widely known Treebanks is the Penn Treebank. The latest version of the Penn Treebank, Treebank-3, contains a heterogeneous collection of texts, including fully parsed articles from the Wall Street Journal as well as parsed versions of the Brown Corpus and the Switchboard Corpus (https://catalog.ldc.upenn.edu/LDC99T42). Treebanks are typically parsed with probabilistic parsers – parsers that are advantageous because they are "able to parse rare or aberrant kinds of language, as well as more regular, run-of-the-mill types of sentence structures" (Leech and Eyes 1997: 35).

Treebanks contain sentences that have been either wholly or partially parsed, and a parser can make use of the already parsed structures in a Treebank to parse newly encountered structures and improve the accuracy of the parser. The example below contains a parsed sentence from the Lancaster Parsed Corpus:

```
A01 2
[S[N a_AT move_NN [Ti[Vi to_TO stop_VB Vi][N \0Mr_NPT
   Gaitskell_NP N][P from_IN [Tg[Vg nominating_VBG Vg][N
   any_DTI more_AP labour_NN life_NN peers_NNS N]Tg]P]Ti]N]
   [V is_BEZ V][Ti[Vi to_TO be_BE made_VBN Vi][P at_IN [N
```

a_AT meeting_NN [Po of_INO [N labour_NN \0MPs_NPTS N]Po]N]P][N tomorrow_NR N]Ti] ._. S]

The first line of the example indicates that this is the second sentence from sample "A01" (the press reportage genre) of the LOB Corpus, sections of which (mainly shorter sentences) are included in the Treebank. Open and closed brackets mark the boundaries of constituents: "[S" marks the opening of the sentence, "S]" the closing; the "[N" preceding *a move* marks the beginning of a noun phrase, "N]" following *to stop* its ending. Other constituent boundaries marked in the sentence include "Ti" (*to*-infinitive clause *to stop Gaitskell from*. . .), "Vi" (non-finite infinitive clause *to stop*), and "Vg" (non-finite *–ing* participle clause *nominating*). Within each of these constituents, every word is assigned a part of speech tag: *a*, for instance, is tagged "AT", indicating it is an article; *move* is tagged "NN", indicating it is a singular common noun; and so forth. Although many Treebanks have been released and are available for linguistic analysis, their primary purpose is to train parsers to increase their accuracy. The Survey of English usage at University College London has as a page on its website (www.ucl.ac.uk/english-usage/projects/ice-gb/compare.htm) that provides a detailed description and comparison of the various Treebanks and parsed corpora that are available.

3.13 Conclusions

The process of collecting and computerizing texts is, as this chapter has demonstrated, a labor-intensive effort. For instance, recording and transcribing spontaneous conversations requires considerable time because individuals need to be recorded and their recordings transcribed. While voice recognition software has made tremendous progress recently, it still works best on monologic speech: an individual speaking slowly into a microphone that is then transferred into written text in a document. Such programs do not work well with dialogic speech.

Written texts, in contrast, are now widely available in digital formats and can easily be incorporated in a corpus after permission has been received to use a given text. While lexically tagging corpora with part-of-speech information can now be done quickly and quite accurately, parsing a corpus is a much more difficult undertaking, since the level of accuracy decreases when identifying structures such as noun phrases or imperative sentences.

The Web has also increased the availability of texts, even certain types of spoken texts. The BYU corpora, for instance, contain different kinds of public speech made available in transcripts that are easily obtainable. While their accuracy cannot be guaranteed, the level of error does not appear to be high. It is also the case that voice recognition software will continue to improve so that in the future it could very well be possible to automate the process of converting speech to text, and thus expedite the inclusion of, for instance, spontaneous dialogues in corpora.

4 Analyzing a Corpus

The process of analyzing a completed corpus is in many respects similar to the process of creating a corpus. Like the corpus compiler, the corpus analyst needs to consider such factors as whether the corpus to be analyzed is lengthy enough for the particular linguistic study being undertaken and whether the samples in the corpus are balanced and representative. The major difference between creating and analyzing a corpus, however, is that while the creator of a corpus has the option of adjusting what is included in the corpus to compensate for any complications that arise during the creation of the corpus, the corpus analyst is confronted with a fixed corpus, and has to decide whether to continue with an analysis if the corpus is not entirely suitable for analysis, or find a new corpus altogether.

This chapter describes the process of analyzing a completed corpus. To illustrate how such analyses are conducted, the chapter opens with a discussion of Former President Donald Trump's usage of language, termed "Trump speak," and how corpora of his Twitter posts, transcribed speeches, and other spontaneous commentary can be used to study his unique uses of language. The discussion in this section illustrates how to (1) frame a research question, (2) select relevant corpora to carry out the analysis, (3) use a concordancing program to locate appropriate examples for analysis, and then (4) explain the results drawing upon relevant research on language usage, particularly theories of politeness.

The remainder of the chapter provides an overview of qualitative and quantitative research methodologies, discussing the differences between the two methodologies and providing examples of each. For instance, many of the corpus-based reference grammars of English, such as Quirk et al.'s *A Comprehensive Grammar of the English Language*, are more qualitative, as they draw upon linguistic corpora for authentic examples to illustrate the many points of English grammar discussed throughout the grammar.

In contrast, other corpus studies are more *quantitative*, subjecting the results of a corpus analysis to, for instance, statistical analyses to

determine whether the particular linguistic differences in corpora under study are significant or not. For instance, Biber (1988 and elsewhere) uses *multi-dimensional [MD] analysis* in much of his research to study whether the differences in the distribution of linguistic features in corpora are statistically significant or not (e.g. whether the first person pronoun *I* is less likely to occur in formal writing than in spontaneous dialogues). Other corpus analyses draw upon *descriptive statistics*, such as the log likelihood ratio, a statistical test, similar to Chi square, to determine whether distributions of linguistic features are statistically significant. Sample analyses are included to illustrate the use of both descriptive statistics and MD analysis.

4.1 *Trump Speak*: Framing a Research Question

To determine exactly what research question one wishes to pursue, it is first of all necessary to review relevant articles or books written on the topic so that the ultimate question selected does more than merely repeat what others have written on the topic.

Although Donald Trump (hereafter DT) was new to the political scene – President of the United States was his first elected office – he was quite well-known prior to becoming President as a businessperson and television celebrity. However, in his relatively short time as president, he has established a very distinct persona. In particular, he flouts on a regular basis the norms of speech that one would expect from someone in his position. Consequently, his style of speaking has drawn considerable interest from corpus linguists.

Clarke and Grieve (2019) conducted a stylistic analysis of Trump's tweets in the Trump Twitter Archive between the years 2009 and 2021. This archive is very comprehensive and contains every Tweet that Trump posted from 2009 to January 8, 2021, when Trump was banned from Twitter and his account was closed. In their analysis, they note, for instance, changes in the length and frequency of tweets over time and instances when Trump was particularly active in criticizing a particular individual. For instance, the tweets increased in frequency when Trump engaged in a lengthy campaign questioning Obama's citizenship (p. 3).

To study Trump's style of communication, Clark and Grieve adapted Biber's notion of multi-dimensional analysis (see Section 4.6) to isolate certain features of Trump's tweeting style. In his work

on register variation, Biber develops a series of what he calls dimensions: general parameters that describe a particular style of communication. For instance, Clark and Grieve's (2019: 18) Dimension 5: *Advisory Style* characterizes tweets in which Trump is giving advice:

> Sorry losers and haters, but my I.Q. is one of the highest -and you all know it! Please don't feel so stupid or insecure, it's not your fault
> (332308211321425920, 2013-05-09, D5: 0.745)

This dimension contains such features as imperative sentences and second person pronouns as a means of showing that Trump is communicating directly with his audience and offering them some kind of advice. This advice can be either positive or in the case above negative and critical.

Other corpus linguists have also written about Trump's style of speaking. For instance, Xueliang Chen, Yuanle Yan, and Jie Hu (2019) conducted a corpus analysis of the use of language by Hillary Clinton and Donald Trump when they were running against each other for president. They considered two research questions in their analysis:

(1) What are the linguistic features of Clinton's and Trumps' speeches, and what themes do they reflect in each candidate's campaign?

(2) What are the differences between Clinton's and Trump's linguistic styles, and what political beliefs do they represent? (p. 14)

To answer these questions, they analyzed two comparable corpora that they created, both containing, respectively, speeches presented by Trump and Clinton as they were running against each other. After analyzing the keywords (words with unexpected high frequencies) in the two corpora they created, they concluded that while Clinton's speeches were more positive in nature with top keywords such as *women's rights*, *social justice*, and *kind*, Trump's keywords were much more negative: *bad*, *illegal*, *disaster*. In fact, two of the top keywords in Trump's speeches were *Hillary* and *Clinton*, reflecting his frequent reference to her in his speeches and the emphasis he put on critiquing her and her policies.

4.2 Selecting Suitable Corpora to Address a Particular Research Question

Because a study of Trump's very often negative use of language is very focused and is a rather narrow topic of research, it

will be necessary to draw upon data from very specific corpora to carry out this study. As a consequence, three corpora were analyzed: the Trump Twitter Archive (as described above), the Web, and a 500,000-word archive of, for instance, Trump's speeches and interviews (www.rev.com/blog/transcript-category/donald-trump-transcripts). Because these are not traditional sources of data, it is worth discussing why they can be considered corpora.

First, corpora typically contain excerpts of texts taken from larger sources. For instance, the Brown Corpus contains 2,000-word samples taken from complete texts (e.g. newspaper articles). But because tweets are relatively short and are not part of any larger text, can they be considered texts themselves? For Halliday and Hasan (1976), a text has unity of structure and unity of texture. While a Tweet may only consist of a limited number of words, it is arguably a complete text, and consequently has unity of structure:

> Biden was asked questions at his so-called Press Conference yesterday where he read the answers from a teleprompter. That means he was given the questions, just like Crooked Hillary. Never have seen this before!
>
> (Trump Twitter Archive, July 1st, 2020)

The tweet above has a topic: Biden's reading his answers to questions from the press on a Teleprompter, and enough sentences to develop this topic so that the entire tweet can stand alone as a coherent unit. Consequently, it has unity of structure. In addition, the tweet has unity of texture: links that tie together the various parts of the tweet. For instance, the tweet opens with an initial reference to Biden. As the tweet develops there are two instances of the pronoun *he* that co-refer to Biden, tying sections of the tweet together and thus creating cohesion and ultimately coherence.

While the Web has increasingly become a corpus used for linguistic analysis, the other two sources of data – an archive of tweets and a collection of Trump's speeches and interviews – are specific to this particular analysis. But as was discussed in Chapter 1 (Section 1.5), there is a range of other types of corpora that can be used for analysis, including multi-purpose corpora, learner corpora, historical corpora, and parallel corpora.

4.3 Extracting Information from a Corpus

There are various ways that grammatical and lexical information can be retrieved from corpora. In the pre-electronic era, such

information had to be manually retrieved. For instance, when Jespersen was looking for authentic examples to illustrate the various grammatical constructions he included in his seven-volume *A Modern English Grammar on Historical Principles*, he, along with a number of student helpers, had to read numerous books and periodicals to obtain authentic examples to illustrate the grammatical points that he was making. Obviously, this involved considerable time and effort.

As sources of data became computerized, concordancing programs were created that allow for various constructions (e.g. words or phrases) to be automatically retrieved from a corpus. For instance, in the analysis of Trump Speak, various examples of usages by Trump were retrieved using either AntConc (described in greater detail later in this section) or a concordancing program specifically designed to retrieve constructions in the Trump Twitter Archive.

Concordancing programs retrieve structures and present them in KWIC (key word in context) concordances. For instance, below is a sample list of examples of Trump's use of the word *loser(s)*. This archive contains a built-in concordancing program that highlights the construction being searched, and also contains a short context in which, in this case, *loser* (158 examples) or *losers* (156 examples) occur. Below are some sample tweets retrieved by this program:

(3) Steve Scully of @cspan had a very bad week. When his name was announced, I said he would not be appropriate because of conflicts. I was right! Then he said he was hacked, he wasn't. I was right again! But his biggest mistake was "confiding" in a lowlife **loser** like the Mooch. Sad! (Oct 16th, 2020 – 8:15:38 AM EST)

(4) How dare failed Presidential Candidate (1% and falling!) @CoryBooker make false charges and statements about me in addressing Judge Barrett. Illegally, never even lived in Newark when he was Mayor. Guy is a total **loser**! I want better Healthcare for far less money, always… (Oct 13th, 2020 – 5:56:41 PM EST)

(5) Joe Biden is a PUPPET of CASTRO-CHAVISTAS like Crazy Bernie, AOC and Castro-lover Karen Bass. Biden is supported by socialist Gustavo Petro, a major **LOSER** and former M-19 guerrilla leader. Biden is weak on socialism and will betray Colombia. I stand with you! (Oct 10th, 2020 – 2:38:59 PM EST)

(6) Because I've beaten him and his very few remaining clients so much, and so badly, that he has become a blathering idiot. He failed with John McCain and will fail again with all others. He is a total **loser**.

@MarshaBlackburn is a Tennessee Star, a highly respected (WINNER! Oct 7th, 2020 – 8:18:13 AM EST)

This search yielded 314 examples of Trump's use of the expression *loser(s)*.

KWIK concordances are useful because they can easily and quickly be created and also provide the context in which search terms occur. Consequently, when it comes time to include examples in an article or presentation, one can easily integrate the examples into the discussion, and obtain frequency information that indicates whether the usage is common or uncommon.

The results of searches can also help in establishing trends in a corpus. For instance, when Trump was running for the Republican nomination for president, he had nicknames for each of the individuals against whom he was running. For instance, Marco Rubio was commonly referred to by Trump as *Little Marco* (109 mentions in the Trump Twitter Archive); Ted Cruz was *Lyin' Ted Cruz* (23 mentions); and Jeb Bush was *Low Energy Jeb (*or variations, e.g. *Jeb low energy*) (4 mentions). During his run for the presidency, Trump frequently referred to Hillary Clinton as *Crooked Hillary* (366 mentions).

A search of these names on the Web yields huge returns. For instance, a search for *Little Marco* yielded 345 million hits (May 16th, 2021). Similarly, large figures could be found for the other candidates as well. However, frequency information from the Web has to be cautiously interpreted because the returns may, for instance, come from webpages with identical content. Still, web examples and frequencies can provide at least a preliminary sense of how frequent the constructions occur – frequencies that can then be double-checked in other sources.

The concordancing program included with the Trump Twitter Archive is only able to retrieve individual strings of words from a corpus of Trump's tweets. However, other concordancing programs, such as AntConc, allow searches of corpora that are directly linked with AntConc. For instance, the concordancing window in Figure 4.1 is based on a search of a corpus containing a file consisting of miscellaneous texts that were loaded directly into the program. It would have been possible to have retrieved all of these forms simply by searching individually for *talk*, *talking*, and *talked*. However, if you wish to study the different forms of a particular word, you simply add the symbol * at the end of the word. The search term **talk*** retrieved *talk*, *talking*, and *talked*.

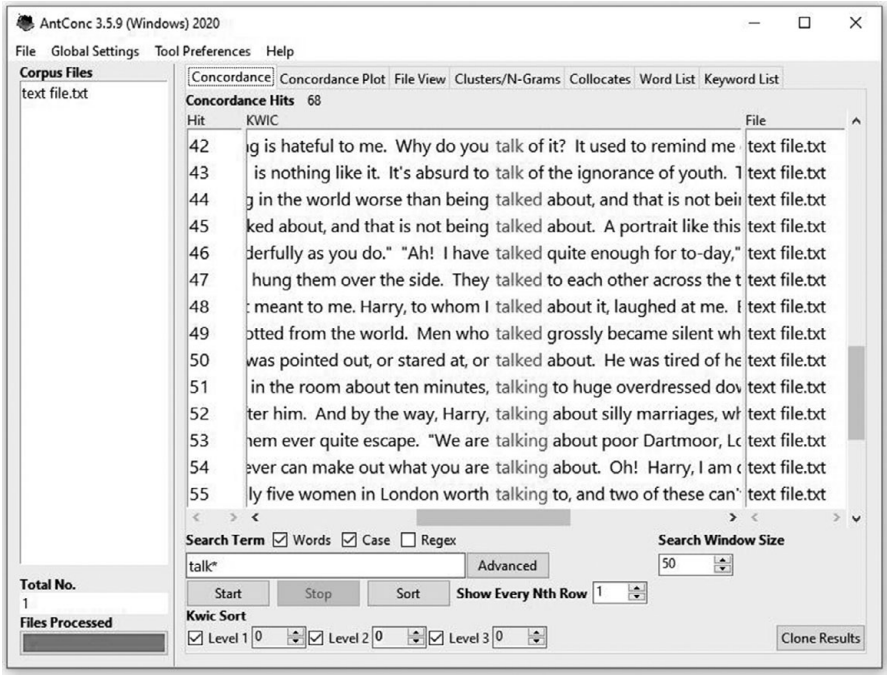

Figure 4.1 Search results for all forms of *talk*

Notice in the table that all of the search terms that were retrieved are highlighted and vertically aligned for easy reading. Also included is a short context containing a span of text that precedes and comes after the search terms. The size of the search window can be widened or narrowed depending upon how much text is desired both before and after the search term.

AntConc, as the top row reveals, can do other kinds of searches. Figure 4.2 displays search results based on the same corpus but searching for clusters and N-grams.

N-grams or clusters are groups of words that co-occur. For instance, the expression *talk about* occurred 6 times in the corpus. Retrieving N-grams in corpora can, for instance, be useful for studying collocations: words that commonly occur together.

The BYU corpora are a set of many different corpora taken from various online resources. The corpora available range from a corpus of American English, the Corpus of Contemporary American English (COCA), to the Coronavirus Corpus, a corpus containing texts retrieved from the Web containing discussions of the Covid-19 virus.

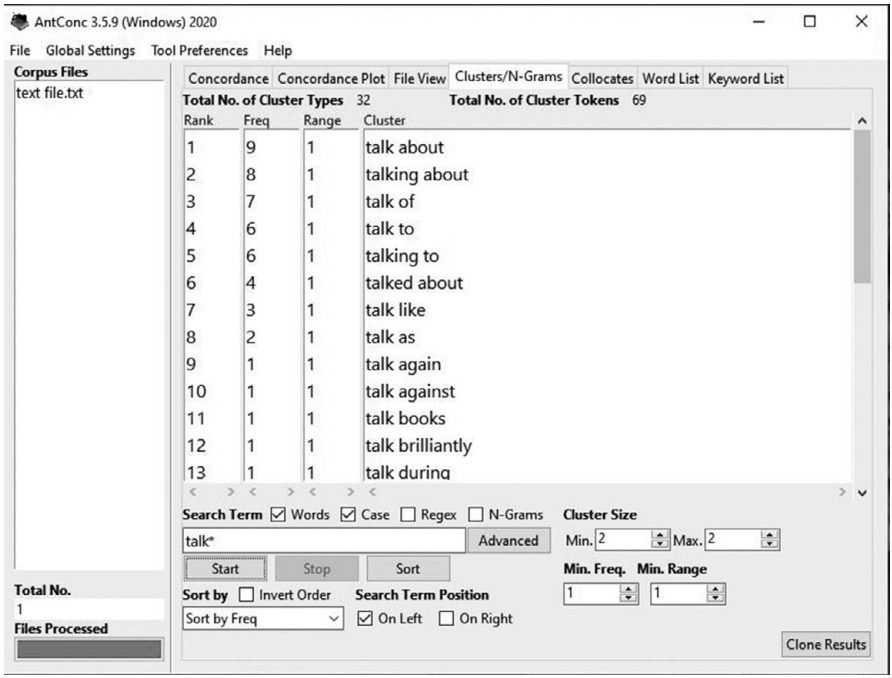

Figure 4.2 Search results for clusters and N-grams

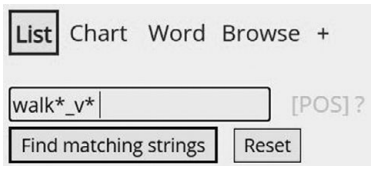

Figure 4.3 Search for all forms of the verb *walk*

All told, 19 corpora are available for online analysis, and all are quite large. For instance, COCA is currently one billion words in length. The size of each corpus constantly changes because all corpora featured on the site are monitor corpora: corpora to which new texts are added on a regular basis.

All BYU corpora can be searched online with a custom-made concordancing program. Figure 4.3 is an example of a search for all forms of the verb *walk*.

The search terms specify that *walk* as a verb be retrieved rather than, for instance, *walk* as a noun. In addition, the symbol * ensures that all forms of the verb *walk* will be retrieved, ignoring forms of the

word *walk* that are nouns. Figure 4.4 contains frequency statistics for four of the more frequently occurring forms that were retrieved.

Because of the size of COCA, each of these forms are displayed in separate KWIK concordances. Therefore, selecting *walk* (which occurred 92,178 times) generates the following KWIC concordance (Figure 4.5), which features the first seven concordance lines that were retrieved:

Another concordancing program, BNCweb (http://corpora.lancs.ac.uk/BNCweb/), allows for web-based searches of the entire British National Corpus, a corpus that contains 100 million words of spoken and written British English. Figure 4.6 contains the first 11 hits of a

☐	CONTEXT ALL FORMS (SAMPLE): 100 200 500	FREQ
☐	WALK	92178
☐	WALKED	81073
☐	WALKING	57989
☐	WALKS	22416

Figure 4.4 Search results for all forms of the verb *walk*

1	i will deem at friends albeit cyber friends you will **walk** lonely within this game and no doubt RL (ICE) (optc)
2	questioning, why do only the male dwarves get to **walk** around without their shirt on? # Gimli: It's true you do
3	'll be here all week. # While trying to **walk** a mile in someone else's shoes and understand the opposing political party's
4	post Gadgets for the Pros, but now I'll **walk** you through the process. # I get a lot of stuff done in
5	all see how in her element she is when we **walk** in my parents' woods. She's pretty much always happy. Here
6	other pack members to these strange funny looking animals that **walk** on 2 legs. # Gil # Alaska # October 8, 11:12 am
7	and love. # Come, Lord Jesus, and **walk** with me along the Way that leads to You. # Thank you,

Figure 4.5 Examples of forms of *walk* retrieved

Your query "*ion" returned 1337927 hits in 3975 different texts (98,313,429 words [4,048 texts]; frequency: 13608.79 instances per million words)

|< << >> >| Show Page: 1 Show KWIC View Show in random order Show e

No	Filename	Hits 1 to 50 Page 1 / 26759
1	A00 2	AIDS (Acquired Immune Deficiency Syndrome) is a **condition** caused by a virus called HIV (Human Immuno Deficiency Virus)
2	A00 3	This virus affects the body's defence system so that it cannot fight **infection**.
3	A00 4	How is **infection** transmitted?
4	A00 11	The medical aspects can be cancer, pneumonia, sudden blindness, dementia, dramatic weight loss or any **combination** of these.
5	A00 12	Often infected people are rejected by family and friends, leaving them to face this chronic **condition** alone.
6	A00 15	10 **million** people worldwide are infected with HIV.
7	A00 18	women are twice as at risk from **infection** as men.
8	A00 24	The World Health **Organisation** projects 40 million infections by the year 2000.
9	A00 24	The World Health Organisation projects 40 **million** infections by the year 2000.
10	A00 25	'We are just at the beginning of the worldwide epidemic and the **situation** is still very unstable.
11	A00 29	ACET — Practical home care, schools **education** and training — 081 840 7879

Figure 4.6 Search results for words ending in the suffix *-ion*

KWIK concordance based on a search of the suffix *ion. The asterisk restricts the search to a variable number of characters occurring before the *-ion*.

The concordancing programs described thus far provide representative examples of the types of searches that such programs allow. However, there are additional concordancing programs as well:

Lancsbox (http://corpora.lancs.ac.uk/lancsbox/index.php) is quite versatile, permitting a number of different searches. It can be used to analyze data either collected by the user, or available in existent corpora. It has multilingual capabilities, and accepts as input data in English as well as other languages. In addition to displaying KWIK concordances and word frequencies, the program can generate collocations as well as special visualizations in concordance windows.

WordSmith (www.lexically.net/wordsmith/) is one of the earlier concordancing programs that was available. It can display KWIK concordances and word lists as well as keywords: words in a particular text that are more frequent than would be expected and thus point to their particular importance in a given text.

The concordancing programs described thus far provide representative examples of the types of searches that such programs allow. However, there are additional programs available on the "Tools for Corpus Linguistics" website (https://corpus-analysis.com/tag/concordancer#list).

While concordancing programs are one way to extract constructions from corpora, there are other ways to do so as well. Gries (2017: 187–8) discusses how to use the R Programing language to retrieve linguistic constructions from corpora. For instance, he describes the process of writing a script (a file containing commands) to retrieve character 3-grams (3 sequences of characters) from a corpus. The advantage of R is that it allows for customizing particular searches.

Concordancing programs are able to locate and identify various sequences of words. However, if a corpus is parsed (i.e. contains larger structures such as phrases, clauses, and sentences that are annotated), it is possible to retrieve through searches much larger structures, such as phrases and clauses.

One lexically tagged and grammatically parsed corpus is ICE-GB, the British component of the International Corpus of English. Each individual word in the corpus is assigned a lexical tag (e.g. noun, verb, preposition, etc.). In addition, phrases and clauses are annotated as well as sentence functions, such as subject, direct object, and adverbial.

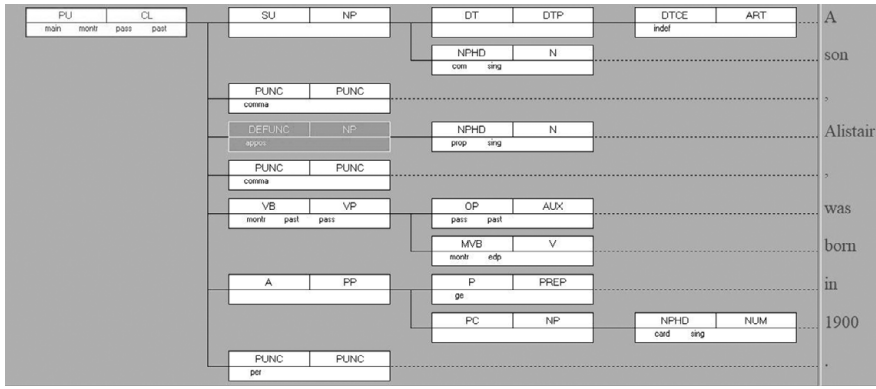

Figure 4.7 Example parsed sentence in ICE-GB

Because ICE-GB is both tagged and parsed, it is possible to retrieve many different types of structures, not just individual words but phrases and clauses as well as other types of constructions, such as appositives. Figure 4.7 contains a parse tree of a sentence that opens with an appositive *A son*, *Alistair*.

All parse trees in ICE-GB contain features, functions, and categories.

In the parse tree in Figure 4.7, the **Feature** "appos" is assigned to the two noun phrases that are in apposition. In such constructions, the second noun phrase either co-refers to the first noun phrase or describes it. The second noun phrase in the apposition, *Alistair*, has two features: it is a singular proper noun. **Functions** describe relationships within clauses or phrases.

For instance, *A son, Alistair* is functioning as "subject" (SU) in the main clause in which it occurs. **Categories** are represented at both the phrase and word level: *A son* is a "noun phrase" (NP), as is *Alistair*.

To find all instances of proper nouns annotated with the feature "appo," ICECUP requires that a "fuzzy tree fragment" (FTF) be constructed (see Figure 4.8); that is, a partial tree structure that can serve as the basis for a search that will retrieve all structures in ICE-GB that fit the description of the construction being searched.

As searches are conducted, it is necessary to collect other types of information that may be relevant for a future paper or article. For instance, any statistical information, such as frequencies, can be included in a spreadsheet. Relevant examples can be saved in a word processing program and include detailed information where the examples originated. Including this information as data is being

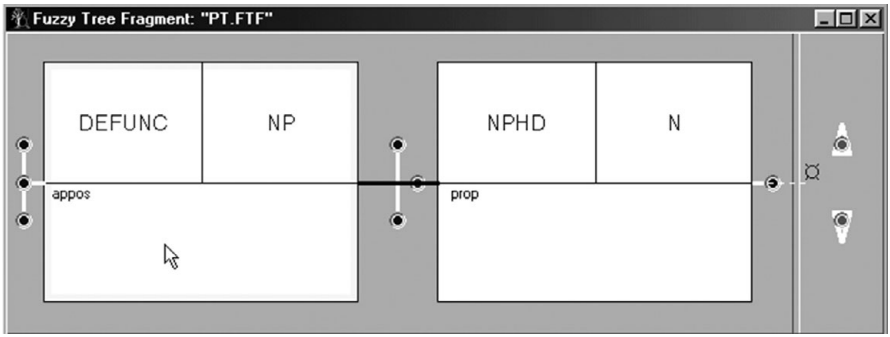

Figure 4.8 Contains an FTF that will find all instances of proper nouns with the feature "appo"

collected can save future time trying to reconstruct where the examples originated.

4.4 Integrating Relevant Linguistic Theories into Corpus Results

A common complaint about corpus analyses is that they often rely heavily on frequency information and statistical analyses, ignoring how this information is connected with relevant linguistic theories. While corpus linguistics has very little to say to those with interests in, for instance, generative grammar, it is of great value to those who are interested in studying pragmatics: how language is used.

One area of pragmatics that has been extensively studied is politeness: how people adjust their manner of speaking to conform to the norms of politeness for the language that they speak. These norms are contextually based: what would be considered impolite in one context may be perfectly acceptable in another context. An important characteristic of Trump Speak is that it often violates the norms of polite speech for both a candidate running for the presidency as well as an elected president. Although theories of politeness in human language are generally conceptualized to explain face-to-face conversations, these theories can be extended to a newer medium, such as Twitter, which is conversational in structure but, unlike interactive human communication, is monologic. For instance, one way that Trump amplifies the impoliteness of his tweets is by assigning particularly negative adjectives to rival candidates' first names. Consider below a series of tweets all containing the expression *Crooked Hillary*:

- Aug 25, 2018 08:11:28 AM The FBI only looked at 3,000 of 675,000 **Crooked Hillary** Clinton Emails.
- Aug 25, 2018 08:05:42 AM Big story out that the FBI ignored tens of thousands of **Crooked Hillary** Emails, many of which are REALLY BAD. Also gave false election info. I feel sure that we will soon be getting to the bottom of all of this corruption. At some point I may have to get involved!
- Aug 22, 2018 07:56:35 PM The only thing that I have done wrong is to win an election that was expected to be won by **Crooked Hillary** Clinton and the Democrats. The problem is, they forgot to campaign in numerous states!
- Aug 19, 2018 06:30:41 AM No Collusion and No Obstruction, except by **Crooked Hillary** and the Democrats. All of the resignations and corruption, yet heavily conflicted Bob Mueller refuses to even look in that direction. What about the Brennan, Comey, McCabe, Strzok lies to Congress, or Crooked's Emails!
- Aug 14, 2018 08:01:50 AM Fired FBI Agent Peter Strzok is a fraud, as is the rigged investigation he started. There was no Collusion or Obstruction with Russia, and everybody, including the Democrats, know it. The only Collusion and Obstruction was by **Crooked Hillary**, the Democrats and the DNC!

Through August 20, 2020, there were 369 instances of **Crooked Hillary** in the archive.

Other derogatory descriptors found in the archive include:

"loser" 324 tweets

Mini Mike Bloomberg is a LOSER who has money but can't debate and has zero presence.

"dummy" 222 tweets and variations on *"dumb"* 194 tweets

Nervous Nancy is an inherently "**dumb**" person.

CNN. Don Lemon is a lightweight - **dumb** as a rock

Mitt Romney, who was one of the **dumbest** and worst candidates in the history of Republican politics

Kirsten Powers is a **dummy**—wasn't she Anthony Weiner's girlfriend?

"*stupid*" 241 tweets

It would be so easy to fix our weak and very **stupid** Democrat inspired immigration laws.

Straighten out The Republican Party of Virginia before it is too late. **Stupid**! RNC

Further similarly negative expressions include "weak" (156); "dope" or "dopey" (117); "moron(s)" (54); and "dishonest" (115).

Most of the examples documented above are insults, a type of speech act that violates many norms of politeness, particularly because Trump's insults of individuals occur in very public forums, such as Twitter, debates, and campaign rallies.

In their discussion of politeness in human language, Brown and Levinson (1987: 60–1) argue that a key component of politeness is the notion of face – "the public self-image that every member wants to claim for himself [or herself]" – and the efforts made by interlocutors to "maintain each other's face." A violation of politeness norms leads to a face threatening act (FTA).

Even a cursory overview of Trump's tweets reveals numerous instances of FTAs – Michael Bloomberg is a loser, Hillary Clinton is crooked, Mitt Romney is the dumbest and worst candidate. Other individuals are stupid, weak, dopey, and dishonest. And while these insults, as noted earlier, would have had greater impact if they were uttered in face-to-face interactions with Trump, the fact that Twitter and other media forums are very public, these insults resonate perhaps more than if they were done privately in a face-to-face conversation.

The frequent co-occurrence of negative descriptors with particular individuals can also result in what is referred to as a negative semantic prosody. For instance, if the expressions "Crooked Hillary" or "Little Marco" occur so frequently, over time, Hillary Clinton will be viewed as crooked and Marco Rubio as diminutive in stature.

One of Leech's (1983) maxims of his politeness principle is also relevant to the impoliteness inherent in Trump's tweets (*O*=other):

- Approbation Maxim: minimize dispraise of *O* [and maximize praise of *O*]

Although this maxim is not universal, it is a given in Western culture that it is better to praise rather than dispraise someone, and if dispraise is given, such speech acts are mitigated. Thus, a teacher might say to a student that his/her paper "has good ideas, but needs some work" rather than "this paper is a disaster, and needs to be completely rewritten." However, in Trump Speak, individuals are dumb, stupid, dopey, and dishonest. Nothing is mitigated.

While much of Trump Speak is highly negative, there are two adjectives that Trump regularly uses that are much more positive: *tremendous* (297) and *incredible* (356). As the frequency counts

demonstrate, these adjectives are two of the most frequently occurring usages in Trump Speak. In addition, they are generally highly positive in contrast to the highly negative connotations of so many of his other usages. While the first example below contains a usage of *tremendous* that is rather neutral in tone, the other usages are quite positive. For instance, he receives "tremendous support" from unions, the win by the US golf team was a "tremendous win." Likewise, in the last example, the Minneapolis police are characterized as "incredible." These usages are noteworthy because all the other previously discussed usages have been predominantly negative in tone.

- As bad as the I.G. Report is for the FBI and others, and it is really bad, remember that I.G. Horowitz was appointed by Obama. There was **tremendous** bias and guilt exposed, so obvious, but Horowitz couldn't get himself to say it. Big credibility loss. Obama knew everything!
- Dec 14, 2019 11:53:55 PM Congratulations to Tiger and the entire U.S. Team on a great comeback and **tremendous** WIN. True Champions!
- Dec 10, 2019 09:32:30 AM America's great USMCA Trade Bill is looking good. It will be the best and most important trade deal ever made by the USA. Good for everybody - Farmers, Manufacturers, Energy, Unions - **tremendous** support. Importantly, we will finally end our Country's worst Trade Deal, NAFTA!
- A GREAT evening in Minneapolis, Minnesota with **incredible** American Patriots. THANK YOU!
- Oct 8, 2019 01:40:40 PM. . .In fact, the "Cops For Trump" T-shirt Web Site CRASHED because of **incredible** volume, but is now back up and running. Proceeds go to the Police Union Charities. See you on Thursday night in Minneapolis!
- Oct 8, 2019 01:40:38 PM Radical Left Dem Mayor of Minneapolis, Jacob Frey, is doing everything possible to stifle Free Speech despite a record sell-out crowd at the Target Center. Presidents Clinton and Obama paid almost nothing! The Minneapolis Police have been **incredible**. . .

There are also two noun phrases that Trump has used repeatedly in his tweets: *witch hunt* (379) and *fake news* (831).

The term *witch hunt* is used by Trump to disparage any investigation of him or one of his friends. For instance, in the tweets below, his friend Roger Stone's conviction of perjury was a "Witch Hunt," as was both Trump's impeachment as well as the Mueller Report, a report that

investigated Russian interference in the 2016 elections and the possible coordination by members of Trump's administration with the Russians.

- Jul 11, 2020 07:24:11 AM Roger Stone was targeted by an illegal **Witch Hunt** that never should have taken place. It is the other side that are criminals, including Biden and Obama, who spied on my campaign - AND GOT CAUGHT!
- Jul 9, 2020 09:38:12 AM The Supreme Court sends case back to Lower Court, arguments to continue. This is all a political prosecution. I won the Mueller **Witch Hunt**, and others, and now I have to keep fighting in a politically corrupt New York. Not fair to this Presidency or Administration!
- Jun 8, 2020 02:14:47 PM. . .Crooked Hillary Clinton in 2016. They are called SUPPRESSION POLLS, and are put out to dampen enthusiasm. Despite 3 ½ years of phony **Witch Hunts**, we are winning, and will close it out on November 3rd!
- The Impeachment Hoax, just a continuation of the **Witch Hunt** which started even before I won the Election, must end quickly. Read the Transcripts, see the Ukrainian President's strong statement, NO PRESSURE - get this done. It is a con game by the Dems to help with the Election!

The term *Fake News*, which occurs quite frequently in his tweets, is typically used to disparage negative news reports against him in the media. For instance, in the second tweet below, he disparages a former government employee who made negative comments about him. In the third tweet, which is a response to media claims that he has not dealt very well with the coronavirus pandemic, he shifts the topic to the Obama administration's handling of the H1N1 Swine Flu:

- Aug 24, 2020 10:25:48 AM Incredible that @CNN & MSDNC aren't covering the Roll Call of States. **Fake News**! This is what the Republican Party is up against. Also, I'd like to hear the remarks of the Delegates from individual States, rather than @FoxNews anchors. Ridiculous!
- Aug 18, 2020 07:30:45 AM Many thousands of people work for our government. With that said, a former DISGRUNTLED EMPLOYEE named Miles Taylor, who I do not know (never heard of him), said he left & is on the open arms **Fake News** circuit. Said to be a real "stiff". They will take anyone against us!
- Aug 18, 2020 06:10:19 AM Looking back into history, the response by the ObamaBiden team to the H1N1 Swine Flu was

considered a weak and pathetic one. Check out the polling, it's really bad. The big difference is that they got a free pass from the Corrupt **Fake News** Media!

- Aug 16, 2020 12:31:20 PM. @FoxNews is not watchable during weekend afternoons. It is worse than **Fake News** @CNN. I strongly suggest turning your dial to @OANN. They do a really "Fair & Balanced" job!
- Aug 15, 2020 03:06:50 PM More **Fake News** !

One final characteristic of Trump Speak has less to do with his repetition of particular vocabulary items and more with his general use of language, particularly his propensity to frequently lie. For instance, as of January 20, 2021 (the last day of his presidency), the Fact Checker Database had documented 30,573 "fake or misleading claims" that Trump has made since becoming President (www.washingtonpost.com/graphics/politics/trump-claims-database/?utm_term=.27babcd5e58c&itid=lk_inline_manual_2&itid=lk_inline_manual_2).

Although Trump's tendency to lie may seem outside the realm of linguistic theory, one of the maxims of Grice's (1989) Cooperative Principle, the Quality Maxim, stipulates that in conversations individuals should "be truthful":

(7) Do not say what you believe to be false

(8) Do not say that for which you lack adequate evidence (Grice 1989: 27)

If one of the maxims of the cooperative principle is violated, a conversational implicature results; that is, an additional meaning not explicitly inherent in the utterance results. For instance, consider the statement below:

"We built the greatest economy in history, not only for our country, but for the world. We were number one, by far."

This claim, made 360 times by Trump, is proven to be false because there is ample economic documentation that many other countries over a span of many years have far exceeded the current state of the economy during Trump's tenure as president. The conversational implicature of Trump's frequent violation of the quality maxim is that, in general, many people do not trust the veracity of just about any claim that he makes.

Although the analysis of Trump speak in this section has included some statistical information – mainly frequency counts of the various

usages that were discussed – the analysis has been primarily qualitative rather than quantitative: it has relied on an extensive analysis of the data rather than a statistical analysis of the frequency counts that are described. In fact, the frequency counts are included mainly to illustrate that the usages discussed occurred frequently enough to demonstrate the rather negative and, in some cases, derogatory nature of Trump Speak. The next section explores in greater detail the differences between qualitative and quantitative corpus analyses.

4.5 Quantitative and Qualitative Analyses

Although quantitative and qualitative approaches to corpus analysis are often viewed as differing ways of analyzing corpora, the difference between the two approaches is not necessarily discrete. For instance, Angouri (2018: 41) proposes what she characterizes as a QUAL/QUAN spectrum, a spectrum recognizing that "while quantitative research is useful towards generalizing research findings. . .Qualitative approaches are particularly valuable in providing in-depth, rich data." In other words, there are not necessarily analyses that are purely qualitative or purely quantitative. Instead, there is a continuum between the two approaches (sometimes referred to as mixed methods). This section of the chapter, then, will focus primarily on analyses that are predominantly on the qualitative end of this spectrum, or the quantitative.

4.5.1 Qualitative Corpus Analysis

Hasko (2020) provides a detailed description of what qualitative corpus analyses involve.

She notes, for instance, that such analyses are "exploratory" and "inductive" and permit "in-depth investigations of authentic language use" (p. 951). To explain particular patterns of usage, qualitative analyses, she continues, can incorporate non-linguistic elements, including various speaker variables, such as age, gender, and economic background. These variables are important for qualitative analyses because they help reveal how non-linguistic elements (e.g. an individual's age or gender) can influence language usage.

In contrast, quantitative analyses rely more heavily on statistical information – information that is very often based on corpora that have been annotated with lexical tags (marking, for instance, nouns or

verbs) and/or syntactically parsed constructions (such as subjects, objects, and adverbials). These constructions can be automatically retrieved and subjected to statistical analyses that determine whether, for instance, the distributions of the constructions in one genre or another (e.g. fiction or press reportage) is statistically significant.

This is not to suggest that there is always an absolute difference between quantitative and qualitative analyses. The two types of analyses can work together to produce an analysis involving what is sometimes termed mixed methods. Thus, one can view the difference between the two types of analyses being on a gradient: a spectrum on which there are varying degrees of quantitative and qualitative analyses.

Qualitative corpus analyses, as Hasko (2020: 952) comments, have a long tradition, dating back to the pre-electronic era. And there is ample evidence to support this claim. For instance, the lexicographer James Murray developed a methodology for collecting authentic usages of language to both determine and illustrate the meanings of words in the first edition of the *Oxford English Dictionary*. Early grammarians such as Otto Jespersen relied extensively on authentic examples from written texts to write their grammars. Meyer (2009: 209) notes that Jespersen's *A Modern English Grammar on Historical Principles* (1909–1949) was based on an analysis of examples taken from one thousand written sources. While many of these sources were literary texts, a range of other genres were represented as well, including literary criticism, history, philosophy, and biography.

The linguist Randolph Quirk created what is now known as the Quirk Corpus, a corpus of spoken and written British English collected between 1955 and 1985. Before they were digitized, citations of examples illustrating various grammatical categories were only available on slips in file drawers housed at the Survey of English Usage (University College London).

But the first computer corpus for the computational analysis of language, as Hasko (2020: 952) comments, was the Brown Corpus, a corpus that paved the way for not only quantitative corpus research but qualitative studies of language as well, as demonstrated by the particular registers included in the Brown Corpus.

The Brown Corpus marked the beginning of the era of computerized corpora that could be used as the basis for conducting empirically based linguistic investigations of English. Although by modern standards it was fairly short (one million words), it set a standard for the design of many subsequent linguistic corpora, not just of English

Table 4.1 *The composition of the Brown Corpus*

Informative Prose	# of Samples: 374
Press: reportage, editorial, reviews	88
Religion	17
Skills and hobbies	36
Popular lore	48
Belles lettres, biography, memoirs, etc.	75
Miscellaneous: government documents, foundation reports, etc.	30
Learned: natural sciences, humanities, social sciences, etc.	80
Imaginative prose	**# of samples: 126**
General fiction	29
Mystery and detective fiction	24
Science fiction	6
Misc.	67

but of other languages as well (cf. Chapter 1 for a more detailed description of the Brown Corpus).

The Brown Corpus contains 2,000-word samples of edited written American English grouped into two general categories: informative prose (374 samples) and imaginative prose (126 samples) (see Table 4.1):

There were several methodological considerations that guided the structure of the corpus. First, to adequately represent the various kinds of written American English, a wide range of different types of written English were selected for inclusion in the corpus. For instance, the bulk of the corpus contains various kinds of informative prose, including press reportage, editorials, and reviews; government documents; differing types of learned writing; learned writing from, for instance, the humanities and social sciences. In addition to informative writing, the corpus contains differing types of imaginative prose, such as general fiction and science fiction. More examples of informative than imaginative writing were included because informative writing is much more common than imaginative writing.

Since neither informative nor imaginative writing is homogeneous, the corpus contained 2,000-word samples from a range of different articles, periodicals, books, and so forth to provide as broad a range as possible of different authors, books, periodicals, and so forth. This sampling procedure was also used to ensure that an extensive range of

different authors, books, and periodicals were represented. Had lengthier samples been used, the style of a particular author or periodical, for instance, would have been over-represented in the corpus and thus have potentially skewed the results of studies done on the corpus.

The Brown Corpus set the standard for how corpora were organized, and as a consequence, was the catalyst for the creation of several additional corpora that replicated its composition. For instance, the London/Oslo/Bergen (LOB) Corpus contains one million words of edited written British English as well as the same genres and length of samples (2,000 words) as the Brown Corpus with only very minor differences: In several cases, the number of samples within a given genre vary: popular lore contains 44 samples, while in the Brown corpus the same genre contains 48 samples (cf. Oostdijk 1991: 37).

The inclusion in the Brown Corpus of many different types of written English made it quite suitable for the qualitative analysis of language usage. A select overview of research conducted on this corpus makes this point very clear. For instance, Elsness (1982) investigated the distribution of *that* versus zero nominal clauses in sixty-four 2,000-word texts taken from sections of the Brown Corpus:

(9) I know *that* he is here.

(10) I know he is here.

Because the inclusion of *that* is optional in such constructions, Elsness was interested in determining the various factors that influenced the retention or deletion of *that*. For instance, he considered such factors as whether the formality of the particular type of writing influenced the choice of either retaining or deleting *that*; whether *that* deletion was less common in formal writing than in informal writing; whether the particular verb influenced use or non-use of *that*; and whether including or not including *that* resulted in a difference in meaning.

To address these research questions, Elsness restricted his analysis to four different registers of the Brown Corpus: Press Reportage; Belles Lettres, Biography, etc.; Learned and scientific Writings; and Fiction: Adventure and Westerns. Because these are very different registers, Elsness would be able to determine, for instance, whether *that*-deletion has a restricted usage: occurring, for instance, less frequently in more formal registers, such as learned and scientific writing, than in less formal registers, such as fiction. Essentially, what Elsness is doing in this article is using frequency counts, which are

quantitative in nature, to support qualitative claims about the usage of *that*-deletion in various genres of English in the Brown Corpus.

The Brown Corpus was also the first corpus to be lexically tagged. Consequently, it can be viewed as ushering in a new era of quantitative analyses of corpora. For instance, one of its earliest applications involved supplying word frequency lists for the American Heritage dictionary (Francis and Kučera 1982). In fact, most contemporary dictionaries are based on large corpora. Such corpora enable the retrieval of numerous examples of individual words in varying contexts, which is crucial for words that overall have lower frequencies. Additionally, large corpora ensure that a sufficient number of authentic examples can be used to illustrate the various meanings of words.

As corpus linguistics developed as a discipline, and the technology used both to create and analyze corpora increased in sophistication, it became easier to create computerized corpora, conduct more automated analyses of them, and also to subject the results of analyses to statistical analyses. As a result, more quantitative analyses of corpora were able to be conducted.

4.5.2 Quantitative Corpus Analysis

This section explores various ways that quantitative analyses of corpora can be conducted. Such analyses involve the application of various statistical tests to determine whether the hypotheses being tested are statistically significant. Because of the complexity of statistical analyses, the discussion in this section will be restricted to two areas:

(11) The first section describes how descriptive statistics, such as Chi square or log likelihood, can be used to determine whether pseudo-titles in various stigmatized constructions such as *Microsoft president Bill Gates* occur with differing frequencies in numerous samples of press reportage taken from newspapers appearing in the various components of ICE.

Pseudo-titles are related to appositives. Thus, in the example above, a full appositive equivalent would be *the president of Microsoft, Bill Gates*. Appositives contain two co-referential noun phrases separated by a comma pause. In pseudo-titles, the first unit acts more like a modifier: no co-referential relationship exists between the two noun phrases, and there is no comma pause separating them. Pseudo-titles also have a very restricted usage, occurring predominantly in press reportage. However, their usage varies, with some newspapers banning them entirely, always requiring full appositives, while other newspapers, often tabloids but broadsheets as well, use them quite frequently.

(12) The second section describes how multi-dimensional analyses in the work of Douglas Biber can be used to study register variation: how various linguistic constructions occur more or less commonly in differing registers, ranging from press reportage to fiction.

4.5.2.1 The Statistical Analysis of Pseudo-Titles

Previous research on pseudo-titles has documented their existence in American, British, and New Zealand press reportage, and demonstrated that because their usage is stigmatized, certain newspapers (particularly in the British press) prohibit their usage. To examine their usage empirically, the press reportage in seven regional varieties of ICE were examined to determine their usage globally. Table 4.2 lists the number of newspapers containing or not containing pseudo-titles in the various regional components of ICE investigated.

As Table 4.2 illustrates, pseudo-titles have spread to all of the regional varieties of English investigated, and it is only in Great Britain that there were many newspapers prohibiting the usage of pseudo-titles.

Although the results displayed in Table 4.2 reveal specific trends in usage, there were some interesting exceptions to these trends. In ICE-USA, after further investigation, it was found that the one US newspaper that did not contain any pseudo-titles, the *Cornell Chronicle*, actually did allow pseudo-titles. It just so happened that the 2,000-word

Table 4.2 *Number of newspapers containing pseudo-titles in various ICE components*

Country	Newspapers w/o Pseudo-Titles	Newspapers w/ Pseudo-Titles	Total
Great Britain	7	8	15
United States	1	19	20
New Zealand	1	11	12
Philippines	0	10	10
Jamaica	0	3	3
East Africa	0	3	3
Singapore	0	2	2
Totals	9 (14%)	56 (86%)	65 (100%)

sample included in ICE did not have any pseudo-titles. This finding reveals an important limitation of corpora: that the samples included within them do not always contain the full range of usages existing in the language, and that it is often necessary to look further than the corpus itself for additional data. In the case of the *Cornell Chronicle*, this meant looking at additional samples of press reportage from the newspaper. In other cases, it may be necessary to supplement corpus findings with "elicitation tests" (cf. Greenbaum 1984 and de Mönnink 1997): tests that ask individuals to identify constructions as acceptable or not. For pseudo-titles, one might give newspaper editors and writers a questionnaire that elicits their attitudes towards pseudo-titles.

Two newspapers whose style manuals prohibited pseudo-title usage actually contained pseudo-titles. The *New York Times* and one British newspaper, *The Guardian*, contained pseudo-titles but only in sports reportage. This suggests that at least in these two newspapers, the prohibition against pseudo-titles sometimes does not extend to less formal types of writing. In addition, the *New York Times* contained in its news reportage instances of so-called borderline cases of pseudo-titles (see above). Some of the British-influenced varieties of English contained a mixture of British and American norms for pseudo-title usage. Example (13) (taken from ICE-East Africa) begins with a pseudo-title, *Lawyer Paul Muite*, but two sentences later contain a corresponding apposition – *a lawyer, Ms Martha Njoka* – that contains features of British English: a title, *Ms*, before the name in the second part of the appositive, and no punctuation marking the title as abbreviated. In American press writing, typically an individual's full name would be given without any title, and if a title were used, it would end in a period (*Ms.*) marking it as an abbreviation.

(13) *Lawyer Paul Muite* and his co-defendants in the LSK contempt suit wound up their case yesterday and accused the Government of manipulating courts through proxies to silence its critics. . .Later in the afternoon, there was a brief drama in court when *a lawyer, Ms Martha Njoka*, was ordered out after she defied the judge's directive to stop talking while another lawyer was addressing the court. (ICE-East Africa)

Exploring a corpus qualitatively allows the analyst to provide descriptive information about the results that cannot be presented strictly quantitatively. But because this kind of discussion is subjective and impressionistic, it is better to devote the bulk of a corpus study to supporting qualitative judgements about a corpus with quantitative information.

4.5.2.2 Using Quantitative Information to Support Qualitative Statements

In conducting a microscopic analysis of data, it is important not to become overwhelmed by the vast amount of statistical information that such a study will be able to generate, but to focus instead on using statistical analysis to confirm or disconfirm the particular hypotheses one has set out to test. In the process of doing this, it is very likely that new and unanticipated findings will be discovered: A preliminary study of pseudo-titles, for instance, led to the discovery that the length of pseudo-titles varied by national variety, a discovery that will be described in detail below.

One of the most common ways to begin testing hypotheses is to use the "cross tabulation" capability found in any statistical package. This capability allows the analyst to arrange the data in particular ways to discover associations between two or more of the variables being focused on in a particular study. In the study of pseudo-titles, each construction was assigned a series of tags associated with six variables, such as the regional variety the construction was found in, and whether the construction was a pseudo-title or a corresponding apposition. To begin investigating how pseudo-titles and corresponding appositives were used in the regional varieties of ICE being studied, a cross tabulation of the variables "country" and "type" was generated. This cross tabulation yields the results displayed in Table 4.3.

The results of the cross tabulation in Table 4.3 yield raw numbers and percentages which suggest various trends. In ICE-USA, Phil, and

Table 4.3 *The frequency of pseudo-titles and corresponding appositives in the national varieties of ICE*

Country	Pseudo-Title	Appositive	Total
USA	59 (54%)	51 (46%)	110 (100%)
Phil	83 (69%)	38 (31%)	121 (100%)
NZ	82 (73%)	31 (27%)	113 (100%)
GB	23 (23%)	78 (77%)	101 (100%)
Total	247	198	445

NZ, more pseudo-titles than corresponding appositives were used, though ICE-Phil and NZ have a greater percentage of pseudo-titles than does ICE-USA. In ICE-GB, just the opposite occurs: more corresponding appositives than pseudo-titles were used, findings reflecting the fact that there is a greater stigma against the pseudo-titles in British press reportage than in the reportage of the other varieties.

When comparing results from different corpora, in this case differing components of ICE, it is very important to compare corpora of similar length. If different corpora of varying length are compared and the results are not "normalized," then the comparisons will be distorted and misleading. For instance, if one were to count and then compare the number of pseudo-titles in one corpus of 40,000 words and another of 50,000 words, the results would be invalid, since a 50,000-word corpus is likely to contain more pseudo-titles than a 40,000-word corpus, simply because it is longer. This may seem like a fairly obvious point, but in conducting comparisons of the many different corpora that now exist, the analyst is likely to encounter corpora of varying length: corpora such as Brown or LOB are one million words in length and contain 2,000-word samples; the London-Lund Corpus is approximately 500,000 words in length and contains 5,000-word samples; and the British National Corpus is 100 million words long and contains samples of varying length. Moreover, often the analyst will wish to compare his or her results with the results of someone else's study, a comparison that is likely to be based on corpora of differing lengths.

To enable comparisons of corpora that differ in length, Biber, Conrad, and Reppen (1998: 263–4) provide a convenient formula for normalizing frequencies. Using this formula, to calculate the number of pseudo-titles occurring per 1,000 words in the four varieties of ICE in Table 4.3, one simply divides the number of pseudo-titles (247) by the length of the corpus in which they occurred (80,000 words) and multiplies this number by 1,000:

$$(247/80{,}000) \times 1{,}000 = 3.0875$$

The choice of norming to 1,000 words is arbitrary, but as larger numbers and corpora are analyzed, it becomes more advisable to norm to a higher figure (e.g. occurrences per 10,000 words).

Although the percentages in Table 4.3 suggest various differences in how pseudo-titles and corresponding appositives are used, without applying any statistical tests there is no way to know whether the

Table 4.4 *Frequency of occurrence of pseudo-titles in the samples from ICE components*

	1	2	3	4	5	6	7	8	9	10	Total
USA	2	3	0	10	16	2	3	15	1	7	59
Phil	15	11	9	6	4	8	6	15	5	4	83
NZ	24	13	4	10	6	5	4	6	10	0	82
GB	0	0	0	0	8	3	3	2	0	7	23

Minimum	***Maximum***	***Average***	**Standard Deviation**	***Kurtosis***	**Skewness**
0	24	6.175	5.514026	–2.97711	1.147916

differences are real or due to chance. Therefore, in addition to considering percentage differences in the data, it is important to apply statistical tests to the results to ensure that any claims made have validity. The most common statistical test for determining whether differences are significant or not is the t-test, or analysis of variance. However, because linguistic data do not typically have normal distributions, it is more desirable to apply what are termed "non-parametric" statistical tests: tests that make no assumptions about whether the data on which they are being applied have a normal or non-normal distribution.

Data that are normally distributed will yield a "bell curve": most cases will be close to the "mean", and the remaining cases will fall off quickly in frequency on either side of the curve. To understand why linguistic data are not normally distributed, it is instructive to examine the occurrence of pseudo-titles in the 40 texts that were examined (cf. Table 4.4), and the various statistical measurements that calculate whether a distribution is normal or not.

As the figures in Table 4.4 indicate, the distribution of pseudo-titles across the 40 samples was quite varied. Many samples contained no pseudo-titles; one sample contained 24. The average number of pseudo-titles per sample was around six. The standard deviation indicates that 68 percent of the pseudo-titles occurring in the samples clustered within about 5.5 points either below or above the average; that is, that 68 percent of the samples contained between one and 11

Table 4.5 *Chi square results for differences in the distribution of pseudo-titles and corresponding appositives in the samples from ICE components*

Statistical Test	***Value***	***Degrees of Freedom***	***Significance Level***
Chi square	65.686	3	$p \ll .000$

pseudo-titles. But the real signs that the data are not normally distributed are the figures for kurtosis and skewness.

If the data were normally distributed, the figures for kurtosis and skewness would be "0" (or at least close to "0"). Kurtosis measures the extent to which a distribution deviates from the normal bell curve: whether the distribution is clustered around a certain point in the middle (positive kurtosis), or whether the distribution is clustered more around the ends of the curve (negative kurtosis). Skewness measures how "asymmetrical" a distribution is: the extent to which more scores are above or below the mean. Both of the scores for kurtosis and skewness are very high: a negative kurtosis of -2.97711 indicates that scores are clustering very far from the mean (the curve is relatively "flat"), and the figure of 1.147916 for skewness indicates that more scores are above the mean than below.

Because most linguistic data behave the way that the data in Table 4.4 do, it is more desirable to apply non-parametric statistical tests to the results, and one of the more commonly applied tests of this nature in linguistics is the Chi square. The Chi square statistic is very well suited to the two-way cross tabulation in Table 4.5: the dependent variable (i.e. the variable that is constant, in this case the "country") is typically put in the left-hand column, and the independent variable (i.e. that variable that changes, in this case the "type": whether the construction is a pseudo-title or corresponding apposition) is in the top-most row. Table 4.5 presents the results of a Chi square analysis of the data in Table 4.5.

In essence, the Chi square test calculates the extent to which the distribution in a given dataset either confirms or disconfirms the "null hypothesis": in this case, whether or not there are differences in the distribution of pseudo-titles and equivalent appositives in the four regional varieties of ICE being compared. To perform this comparison, the Chi square test compares "observed" frequencies in a given dataset with "expected" frequencies (i.e. the frequencies one would expect to

find if there were no differences in the distribution of the data). The higher the Chi square value, the more significant the differences are.

The application of the Chi square test to the frequencies in Table 4.5 yielded a value of 65.686. To accurately interpret this number, one first of all needs to know the "degrees of freedom" in a given dataset (i.e. the number of data points that may vary). Since Table 4.5 contains four rows and two columns, the degrees of freedom can be calculated using the formula below:

$$(4-1) \times (2-1) = 3 \text{ degrees of freedom}$$

With three degrees of freedom, the Chi square value of 65.686 is significant at less than the .000 level.

While it is generally accepted that any level below .05 indicates statistical significance, it is quite common for more stringent significance levels to be employed (e.g. $p \leqq .001$). Because the significance level for the data in Table 4.5 are considerably below either of these levels, it can be safely and reliably assumed that there are highly significant differences in the distributions of pseudo-titles and appositives across the four varieties of English represented in the table.

The Chi square test applied to the data in Table 4.5 simply suggests that there are differences in the use of pseudo-titles in the four national varieties of English being investigated. The Chi square test says nothing about differences between the individual varieties (e.g. whether ICE-USA differs fromICE-NZ). To be more precise about how the individual varieties differ from one another, it is necessary to compare the individual varieties themselves in a series of 2×2 Chi square tables. However, in examining a single dataset as exhaustively as this, it is important to adjust the level that must be reached for statistical significance because, as Sigley (1997: 231) observes, "If ... many tests are performed on the same data, there is a risk of obtaining spuriously significant results." This adjustment can be made using the Bonferroni correction, which determines the appropriate significance level by dividing the level of significance used in a given study by the number of different statistical tests applied to the dataset. The Bonferroni-corrected critical value for the ICE data being examined is given below and is based on the fact that to compare all four ICE components individually, six differenent Chi square tests will have to be performed:

.05	/	6	= .0083
Significance Level		# of Tests Performed	Corrected Value

Table 4.6 contains the results of the comparison, from most significant differences down to least significant differences.

Table 4.6 *Comparison of the distribution of pseudo-titles and corresponding appositives in individual ICE components*

Countries	Statistical Test	Degrees of Freedom	Value	Significance Level
NZ and GB	Chi square	1	50.938	$p \gg 0.0001$
Phil and GB	Chi square	1	44.511	$p \gg 0.0001$
US and GB	Chi square	1	19.832	$p \gg 0.0001$
US and NZ	Chi square	1	7.796	$p = .005$
US and Phil	Chi square	1	4.830	$p = .028$ (non-sig.)
Phil and NZ	Chi square	1	.273	$p = .601$ (non-sig.)

The results in Table 4.6 illustrate some notable differences in the use of pseudo-titles and equivalent appositives in the various national varieties. First of all, the use of these structures in ICE-GB is highly different from the other varieties: the levels of significance are very high and reflect the deeply ingrained stigma against the use of pseudo-titles in British press reportage, a stigma that does not exist in the other varieties. Second, even though pseudo-titles may have originated in American press reportage, their use is more widespread in ICE-NZ and ICE-Phil, though with the Bonferroni correction the values are just below the level of significance to indicate a difference between ICE-USA and ICE-Phil. Finally, there were no significant differences between ICE-NZ and ICE-Phil. These results indicate that pseudo-title usage is widespread, even in British-influenced varieties such as New Zealand English, and that there is a tendency for pseudo-titles to be used more widely than equivalent appositives in those varieties other than British English into which they have been transplanted from American English.

While the Chi square statistic is a very useful statistic for evaluating corpus data, it does have its limitations. If the analyst is dealing with fairly small numbers resulting in either empty cells or cells with low frequencies, then the reliability of Chi square is reduced. Table 4.7 lists the correspondence relationships for appositives in the four varieties of English examined.

Three of the cells in the category of "Total Equivalence" contain fewer than five occurrences, making the Chi square statistic invalid for the data in Table 4.7. One way around this problem is to combine variables in a principled manner to increase the frequency for a given cell and thus make the results of the Chi square statistic more valid. One reason for recording the particular correspondence relationship

Table 4.7 *Correspondence relationships for appositives in the samples from ICE components*

Country	Total Equiv.	Det. Deletion	Partial Equivalence	Total
USA	1 (2%)	14 (28%)	36 (71%)	51 (100%)
Phil	1 (2.6%)	8 (21%)	29 (76%)	38 (100%)
NZ	0 (0%)	13 (23%)	18 (58%)	31 (100%)
GB	8 (10%)	22 (28%)	48 (62%)	78 (100%)
Total	10 (5%)	57 (29%)	131 (66%)	198 (100%)

Table 4.8 *Correspondence relationships for appositives in the samples from ICE components (with combined cells)*

Country	Total Equiv./ Det. Deletion	Partial Equivalence	Total
USA	15 (29%)	36 (71%)	51 (100%)
Phil	9 (24%)	29 (76%)	38 (100%)
NZ	13 (23%)	18 (58%)	31 (100%)
GB	30 (39%)	48 (62%)	78 (100%)
Total	67 (34%)	131 (66%)	198 (100%)
Statistical Test	***Value***	***Degrees of Freedom***	***Significance Level***
Chi square	3.849	3	p = .278

for an appositive was to study the stylistic relationship between pseudo-titles and various types of equivalent appositives: to determine, for instance, whether a newspaper prohibiting pseudo-titles relied more heavily than those newspapers allowing pseudo-titles on appositives related to pseudo-titles by either determiner deletion (e.g. *the acting director, Georgette Smith → acting director Georgette Smith*) or total equivalence (*Georgette Smith*, *acting director → acting director Georgette Smith*). Because these two correspondence relationships indicate similar stylistic choices, it is justifiable to combine the results for both choices to increase the frequencies and make the Chi square test for the data more valid.

Table 4.8 contains the combined results for the categories of "total equivalence" and "determiner deletion." This results in cells with high enough frequencies to make the Chi square test valid.

The results indicate, however, that there was really no difference between the four varieties in terms of the correspondence relationships that they exhibited: the Chi square value (3.849) is relatively low and as a result the significance level (.278) is above the level necessary for statistical significance.

It was expected that ICE-GB would contain more instances of appositives exhibiting either total equivalence or determiner deletion, since in general British newspapers do not favor pseudo-titles and would therefore favor alternative appositive constructions. And indeed, the newspapers in ICE-GB did contain more instances. But the increased frequencies are merely a consequence of the fact that, in general, the newspapers in ICE-GB contained more appositives than the other varieties. Each variety followed a similar trend and contained fewer appositives related by total equivalence or determiner deletion and more related by partial equivalence. These findings call into question Bell's (1988) notion of determiner deletion, since overall, such constructions were not that common and whether a newspaper allowed or disallowed pseudo-titles had little effect on the occurrence of appositives related by determiner deletion. Particular correspondence relations determined whether the appositive contained a genitive noun phrase or some kind of post-modification, structures that led to a partial correspondence with a pseudo-title and that occurred very commonly in all varieties.

While combining values for variables can increase cell values, often such a strategy does not succeed simply because so few constructions occur in a particular category. In such cases, it is necessary to select a different statistical test to evaluate the results. To record the length of a pseudo-title or appositive, the original coding system had six values: one word in length, two words, three words, four words, five words, and six or more words. It turned out that this coding scheme was far too delicate and made distinctions that simply did not exist in the data: many cells simply had too low a frequency to apply the Chi square test. And combining categories, as is done in Table 4.9, still resulted in two cells with frequencies lower than five, making the Chi square results for this dataset invalid.

In cases like this, it is necessary to apply a different statistical test: the log-likelihood (or G^2) test. Dunning (1993: 65–6) has argued that, in general, this test is better than the Chi square test because it can be applied to "very much smaller volumes of text. . .[and enable] comparisons to be made between the significance of the occurrences of both rare and common phenomenon." Dunning (1993: 62–3) notes that the Chi square test was designed to work with larger datasets that have

Table 4.9 *The length of pseudo-titles in the various components of ICE*

Country	1–4 Words	5 or More Words	Total
USA	57 (97%)	2 (3%)	59 (100%)
Phil	71 (86%)	12 (15%)	83 (100%)
NZ	66 (81%)	16 (20%)	82 (100%)
GB	23 (100%)	0 (0%)	23 (100%)
Total	217 (88%)	30 (12%)	247 (100%)
Statistical Test	***Value***	***Degrees of Freedom***	***Significance Level***
Chi square	*12.005*	*3*	*p = .007*
Likelihood Ratio	15.688	3	p = .001

items that are more evenly distributed, not with corpora containing what he terms "rare events" (e.g. two instances in ICE-USA of pseudo-titles lengthier than five words). Applied to the data in Table 4.9, the log-likelihood test (termed the "likelihood ratio" in SPSS parlance) confirmed that the length of pseudo-titles varied by variety.

The results of the log-likelihood test point to a clear trend in Table 4.9: that lengthier pseudo-titles occur more frequently in ICE-Phil and NZ than in ICE-USA and GB In fact, ICE-GB had no pseudo-titles lengthier than five words, and ICE-USA had only two instances. These findings are reflected in the examples in (14) and (15), which contain pseudo-titles lengthier than five words that occurred predominantly in newspapers in ICE-PHIL and ICE-NZ.

(14a) Salamat and Presidential Adviser on Flagship Projects in Mindanao Robert Aventajado (ICE-Philippines)
(b) Time Magazine Asia bureau chief Sandra Burton (ICE-Philippines)
(c) Marikina Metropolitan Trial Court judge Alex Ruiz (ICE-Philippines)
(d) MILF Vice Chairman for Political Affairs Jadji Murad (ICE-Philippines)
(e) Autonomous Region of Muslime Mindanao police chief Damming Unga (ICE-Philippines)

(15a) Oil and Gas planning and development manager Roger O'Brien (ICE-NZ)
(b) New Plymouth Fire Service's deputy chief fire officer Graeme Moody (ICE-NZ)

Table 4.10 *The length of appositives in the various components of ICE*

Country	1–4 Words	5 or More Words	Total
USA	22 (43%)	29 (57%)	51 (100%)
Phil	14 (37%)	24 (63%)	38 (100%)
NZ	14 (45%)	17 (55%)	31 (100%)
GB	32 (41%)	46 (59%)	78 (100%)
Total	82 (41%)	116 (59%)	198 (100%)
Statistical Test	***Value***	***Degrees of Freedom***	***Significance Level***
Chi square	*.574*	*3*	$p = .902$

(c) corporate planning and public affairs executive director Graeme Wilson (ICE-NZ)
(d) Federated Gisborne-Wairoa provincial president Richard Harris (ICE-NZ)
(e) Wesley and former New Zealand coach Chris Grinter (ICE-NZ)

The pseudo-title is a relatively new and evolving structure in English. Therefore, it is to be expected that its usage will show variation, in this case in the length of pseudo-titles in the various components of ICE under investigation. The appositive, on the other hand, is a well-established construction in English, and if the length of appositives is considered, there were no differences between the varieties, as is illustrated in Table 4.10.

Table 4.10 demonstrates that it is more normal for appositives to be lengthier, and that while ICE-GB has more appositives than the other varieties, the proportion of appositives of varying lengths is similar to the other varieties.

One reason for the general difference in length of appositives and pseudo-titles is that there is a complex interaction between the form of a given pseudo-title or appositive and its length. In other words, three variables are interacting: "type" (pseudo-title or appositive), "form" (simple noun phrase, genitive noun phrase, noun phrase with post-modification), and "length" (1–4 words or 5 words or more). Table 4.11 provides a cross tabulation of all of these variables.

A Chi square analysis of the trends in Table 4.11 would be invalid not only because some of the cells have values lower than five but because the Chi square test cannot pinpoint specifically which variables are interacting. To determine what the interactions are, it is more appropriate to conduct a loglinear analysis of the results.

Table 4.11 *The form and length of pseudo-titles and corresponding appositives*

Type	*Form*	*1–4 Words*	*5 or More Words*	*Total*
PT	Simple NP	216 (90%)	23 (10%)	239 (100%)
	Gen. NP	0 (0%)	0 (0%)	0 (0%)
	Post. Mod.	1 (13%)	7 (87%)	8 (100%)
Total		217 (88%)	30 (12%)	247 (100%)
Appo	Simple NP	52 (84%)	10 (16%)	62 (100%)
	Gen. NP	18 (67%)	9 (33%)	27 (100%)
	Post. Mod.	12 (11%)	97 (89%)	109 (100%)
Total		82 (41%)	116 (59%)	198 (100%)

A loglinear analysis considers interactions between variables: whether, for instance, there is an interaction between "type," "form," and "length"; between "type" and "form"; between "form" and "length"; and so forth. In setting up a loglinear analysis, one can either investigate a pre-determined set of associations (i.e. only those associations that the analyst thinks exist in the data), or base the analysis on a "saturated model"; that is, a model that considers every possible interaction the variables would allow. The drawback of a saturated model, as Oakes (1998: 38) notes, is that because it "includes all the variables and interactions required to account for the original data, there is a danger that we will select a model that is 'too good'. . .[and that finds] spurious relationships." That is, when all interactions are considered, it is likely that significant interactions between some interactions will be coincidental. Thus, it is important to find linguistic motivations for any significant associations that are found.

Because only three variables were being compared, it was decided to use a saturated model to investigate associations. This model generated the following potential associations:

a. type*form*length
b. type*form
c. type*length
d. form*length
e. form
f. type
g. length

Table 4.12 *Associations between various variables*

K	Degrees of Freedom	Likelihood Ratio	Probability	Chi Square	Probability
3	2	.155	.9254	.145	.9300
2	7	488.010	.0000	630.459	.0000
1	11	825.593	.0000	1172.897	.0000

Table 4.13 *Strongest associations between variables*

	Degrees of Freedom	Likelihood Ratio	Probability
Type*form	2	246.965	.0000
Length*form	2	239.067	.0000

Likelihood ratio and Chi square tests were conducted to determine whether there was a significant association between all three variables (a), and between all possible combinations of two-way interactions (b–d). In addition, the variables were analyzed individually to determine the extent to which they affected the three- and two-way associations in 16a–d. The results are presented in Table 4.12.

The first line in Table 4.12 demonstrates that there were no associations between the three variables: both the likelihood ratio and Chi square scores had probabilities where $p > .05$. On the other hand, there were significant associations between the two-way and one-way variables.

To determine which of these associations were strongest, a procedure called "backward elimination" was applied to the results. This procedure works in a step-by-step manner, at each stage removing from the analysis an association that is least strong and then testing the remaining associations to see which is strongest. This procedure produced the two associations in Table 4.13 as being the strongest of all the associations tested.

Interpreted in conjunction with the frequency distributions in Table 4.12, the results in Table 4.13 suggest that while appositives are quite diverse in their linguistic form, pseudo-titles are not. Even though a pseudo-title and corresponding appositive have roughly the same meaning, a pseudo-title is mainly restricted to being a simple noun phrase that is, in turn, relatively short in length. In contrast, the unit of an appositive corresponding to a pseudo-title can be not just a

simple noun phrase but a genitive noun phrase or a noun phrase with post-modification as well.

These linguistic differences are largely a consequence of the fact that the structure of a pseudo-title is subject to the principle of "end-weight" (Quirk et al. 1985: 1361–2). This principle stipulates that heavier constituents are best placed at the end of a structure, rather than at the beginning of it. A pseudo-title will always come at the start of the noun phrase in which it occurs. The lengthier and more complex the pseudo-title, the more unbalanced the noun phrase will become. Therefore, pseudo-titles typically have forms (e.g. simple noun phrases) that are short and non-complex structures, though as Table 4.10 illustrated, usage does vary by national variety. In contrast, an appositive consists of two units, one of which corresponds to a pseudo-title. Because this unit is independent of the proper noun to which it is related – in speech it occupies a separate tone unit, in writing it is separated by a comma from the proper noun to which it is in apposition – it is not subject to the end-weight principle. Consequently, the unit of an appositive corresponding to a pseudo-title has more forms of varying lengths.

The loglinear analysis applied to the data in Table 4.13 is very similar to the logistic regression models used in the GoldVarb variable rule program. GoldVarb (and its pc equivalent Varbrul) has been widely used in sociolinguistics to test the interaction of sociolinguistic variables. For instance, Tagliamonte and Lawrence (2000) used GoldVarb to examine which of seven linguistic variables favored the use of three linguistic forms to express the habitual past: a simple past tense verb, *used to*, or *would*. Tagliamonte and Lawrence (2000: 336) found, for instance, that the type of subject used in a clause significantly affected the choice of verb form: the simple past was used if the subject was a second person pronoun, *used to* was used if the subject was a first person pronoun, and *would* was used if the subject were a noun phrase with a noun or third person pronoun as head.

Although the GoldVarb program was specifically designed for linguists working with variable rules, most corpus linguists can usefully apply the many differing statistical tests provided by any of the commonly available statistical programs. While the Chi square test is one of the more common tests used with linguistic data, as Oakes' (1998: 1–51) survey of statistical tests for corpus data demonstrates, there are a range of other tests as well. There is also a successor to GoldVarb, Rbrul (www.danielezrajohnson.com/rbrul.html), that has similar capabilities but a more user-friendly interface and that runs more quickly on a computer.

4.6 Multi-dimensional Analysis

In his 1988 book, *Variation across Speech and Writing*, Douglas Biber introduced the notion of multi-dimensional analysis and how it could be used to identify significant linguistic differences between spoken and written language. He critiques traditional linguistic beliefs concerning the differences between speech and writing, such as the claim "that speech is more elaborated and complex than writing" (p. 5). Instead, he argues that it is more informative to view the differences in terms of a series of dimensions that he proposes. The purpose of these dimensions is to demonstrate empirically that "neither speech nor writing is primary; that they are rather different systems, both deserving careful analysis" (p. 7).

Below is a list of the six dimensions that Biber proposed, dimensions that are based on a "methodology to empirically analyze the ways in which linguistic features co-occur in texts and the ways in which registers vary with respect to those co-occurrence patterns" (Biber 2019: 12):

Dimension 1: Involved vs. Informational Discourse
Dimension 2: Narrative vs. Non-narrative Concerns
Dimension 3: Situation-Dependent vs. Explicit Reference
Dimension 4: Overt Expression of Persuasion
Dimension 5: Abstract vs. Non-Abstract Information
Dimension 6: On-Line Informational Elaboration (Biber 1988: 122)

Biber (1988: 66–71) developed these dimensions based on data taken from the Lancaster-Oslo/Bergen (LOB) Corpus, a one million-word corpus containing 2,000-word samples representing differing varieties of written British English (see description above); the London-Lund corpus of spoken British English, which contains various types of spoken English, ranging from private conversations to telephone calls; and personal and professional letters that Biber himself supplied.

To understand how each of the Dimensions work, consider how the registers that are associated with Dimension 1 illustrate the spectrum between those registers that are more highly involved (i.e. interactional) versus those that are more informational (i.e. focused primarily on content):

Registers that are more involved:
telephone conversations, face-to-face conversations, personal letters, spontaneous speeches and interviews

Registers that are more informational:
Biographies, press reviews, academic prose, press reportage, official documents

As examples of the differing usage patterns in two registers from Dimension 1, consider the samples of speech and writing below: one a face-to-face conversation, in which the interactions between speakers are "involved" and "interactional," as opposed to a sample of academic prose, which is primarily "informational":

Spontaneous Conversation (excerpted from the Santa Barbara Corpus of Spoken American English) (www.linguistics.ucsb.edu/sites/secure.lsit.ucsb.edu.ling.d7/files/sitefiles/ research/SBC/S)BC011.trn

JAMIE: How [can you teach a three-year-old to] ta=p [2dance2].
HAROLD: [I can't imagine teaching a] –
[2@Yeah2],
really.
JAMIE: . . . (H)=
MILES: . . . Who suggested this to em.
HAROLD: I have no idea.

It was probably my= . . . sister-in-law's idea because,
. . . I think they saw= . . . that movie.
JAMIE: . . . Tap?

[X] [2X2] –

HAROLD: [What] [2was the2],
MILES: [2<X They had X>2] –
HAROLD: the movie with that . . . really hot tap danc[er].
JAMIE: [Oh] that ki=d.
MILES: . . . He was actually here two weeks ago,

and [I missed him].
JAMIE: [at the .. at] the ja=zz .. t[2ap thing or whatever2].
HAROLD: [2Was he a little kid2]?

(www.linguistics.ucsb.edu/sites/secure.lsit.ucsb.edu.ling.d7/files/sitefiles/research/SBC/S)

Academic Writing (excerpted from the written section of the American component of the International Corpus of English)

<p><#>The risk assessment tools being applied to the acquired immunodeficiency syndrome (AIDS) epidemic are those that were developed for chronic diseases. <#>However, the usual formulation of risk assessment parameters, such as odds ratios, rate ratios, relative risks, and risk differences, which are so useful in chronic disease epidemiology, do not provide stable assessments of risk for factors that affect contagion. <#>Because with contagious diseases the outcome in one study subject is related to changes in risk for other study subjects, risk assessment at the population level does not correspond to a

> summation of risks at the individual level as it does with chronic diseases. <#>To relate risk assessments at the individual and population levels, knowledge of contact patterns is essential. <#>The purposes of this paper are 1) to demonstrate the lack of stability of chronic disease risk measures with contagious diseases, 2) to demonstrate how risk assessment for contagious diseases depends upon assessment of contact patterns even when contact patterns do not cause appreciable differences in the overall epidemic pattern, and 3) to present a new formulation for the action of one important determinant of contact patterns in sexually transmitted diseases, namely biased selection of partners from the potential partners encountered. <#>This new formulation supersedes our previous selective mixing formulation (1).</p>
>
> (www.ice-corpora.uzh.ch/en/joinice/Teams/iceusa.html)

Before examining the two samples of speech and writing, it is worth noting that while spontaneous conversations share many similarities, there are considerable differences between different types of academic writing. For instance, there are, as Biber and Gray (2016: 4) note, "striking grammatical differences between humanities and science writing." The particular example of scientific writing included above trended more towards the science writing spectrum of academic writing. It is therefore quite different in structure than, for instance, a literary critique/discussion of a Jane Austen novel.

Table 4.14 contains select words taken from the two samples of writing and speech. Each of these words was searched for in COCA. Currently, this corpus is a billion words in length and contains

Table 4.14 *Occurrences per million words in spontaneous conversation and academic writing*

Word	*Spontaneous Conversation*	*Academic Writing*
Written text		
Assessment	18.54	240.77
Parameters	2.85	54.48
Chronic	9.12	54.38
Ratios	0.60	24.43
Summation	0.70	2.66
Spoken text		
You	17,722.68	926.68
I think	2,200.45	69.50
Imagine	98.86	40.45
Probably	388.34	123.48
Because	2,267.98	1,060.12

samples of different kinds of American English, including academic writing as well as various kinds of spontaneous speech: no private conversations between individuals but other types of spoken English collected from TV and radio programs. Because of the size of this corpus, it is possible to retrieve significant numbers of individual lexical items and the registers in which they occur.

Two of the words taken from the written text, *assessment* and *summation*, are nominalizations: nouns derived from verbs through the addition of the suffixes *–ment* or *–ion*. Because of the abstract nature of nominalizations, they predominate in writing, particularly academic writing, rather than speech. The frequencies bear out this tendency. The three other words – *parameters*, *chronic*, and *ratios* – are also quite specialized and as a consequence occurred more frequently in writing than in speech.

This is not to suggest that these words are never appropriate in spoken contexts. For instance, a search of *chronic* in the spoken sections of COCA returned many examples from news broadcasts:

> Studies suggest our bodies can fight back and repair damage caused by low level, yet chronic, exposure to radiation
>
> (PBS Newshour)

Thus, the use of *chronic* here is more of a report back to a scientific study on radiation rather than part of, for instance, a casual conversation.

The words in the table from the spoken texts also predominate in speech, but their usages, though less frequent, do occur in the written texts as well. Because the pronoun *you* is clearly interactional, involving more than one speaker, it overwhelmingly predominates in speech.

> Well, first of all, I am a member of the Democratic leadership. I've been in the Democratic caucus. But this is what I will also say, if **you** look at polling in this country, what **you** find is that a whole lot of people are dissatisfied with both the Democratic and Republican parties.

However, *you* can also occur in written texts. In the example below, *you* is used because the two sentences occur in an instructional context. Thus, *you* is used in two imperative sentences.

> Write **your** sentences as quickly and as clearly as possible. Make sure **you** complete four sentences.

Before a multi-dimensional analysis can be performed, a corpus needs to be both lexically tagged and then analyzed by a particular

statistical method termed factor analysis. As was noted in Chapter 3, a corpus that is lexically tagged contains part-of-speech information for each word in the corpus. For instance, in the very simple sentence *I like pizza*, the pronoun *I* would be tagged as a first person pronoun; the verb *like* would be as a present tense lexical verb; and the noun *pizza* as a singular common noun. Of course, different tagging programs will have different terminology and provide varying levels of detail about the items being tagged. Lexical tagging is crucial because it will enable the factor analysis program to determine where the various parts of speech occur: e.g. first person pronouns in more interactive texts; passive verbs in more informational texts.

A tagging program used to conduct multi-dimensional analyses is the Biber tagger (Biber 1988). While this tagger can assign part-of-speech tags, as other tagging programs can, it has further capabilities. As Gray (2019: 46) notes, it can also annotate "additional semantic and syntactic features." For instance, traditional tagging programs identify basic features of nouns, such as whether they are proper versus common nouns, or singular versus plural. However, the Biber tagger can assign what are called "secondary tags" (Gray 2019: 47) identifying which nouns are, for instance, nominalizations (e.g *ratification* in the noun phrase *the ratification of the treaty*). Different types of adjectives can be tagged as well, such as whether they occur in the attributive position in a clause (i.e. before the noun that is modified, as in *the* ***lucky*** *person*) or have the form of a past participle (***running*** *water*) (Gray 2019: 48). Additional tags such as these can be very helpful in carrying out multi-dimensional analyses because, for instance, nominalizations, as Biber (1988: 240) notes, occur most commonly in "highly informational genres such as academic prose, official documents, and professional letters."

Lexically tagged texts serve as input for a multi-dimensional analysis, an analysis that Biber describes as:

> a methodology to empirically analyze the ways in which linguistic features co-occur in texts in the ways in which registers vary with respect to those co-occurrence patterns (Biber 2019: 12).

To discover these patterns, it is necessary to subject the data to a factor analysis, a statistical program that is able to isolate both positive and negative correlations between variables (Brezina 2018: 164). Table 4.15, from Biber (1988: 79), illustrates the positive and negative correlations with the linguistic constructions discussed earlier in the chapter.

Table 4.15 *Positive and negative correlations*

	First person pronoun	Questions	Passives	Nominalizations
First person pronouns	1.00			
Questions	.85	1.00		
Passives	–.15	–.21	1.00	
Nominalization	.08	–.17	.90	1.00

Note, for instance, how first person pronouns correlate positively with questions, but negatively with passives and have a very weak positive correlation with nominalizations. Questions, in turn, correlate negatively with passives and nominalizations. These correlations match the distributions of these constructions in the registers of conversation and academic writing discussed earlier in this section.

Since the publication of Biber (1988), there has been a significant amount of research in many areas of linguistics, both theoretical and practical, applying multi-dimensional analyses. *The Longman Grammar of Spoken and Written English* (Biber et al. 1999) is a reference grammar of English focusing on structures occurring in four registers: fiction, academic prose, journalistic language, and conversation. The goal in this reference grammar is not just to describe various types of linguistic structures that occur in English but to also demonstrate that the structures have varying uses in differing registers.

Biber and Gray (2016) explore the notion of grammatical complexity within the context of academic writing. In one section of the book (pp. 87–122), they provide a number of case studies to demonstrate how different academic writing is from other discourse types. For instance, in a comparison of grammatical complexity in academic research writing with casual conversations, they note the common assumption that academic writing is considered "structurally complex" (p. 88) because it contains lengthier sentences as well as syntactically complex clauses. In contrast, conversations are assumed to contain much simpler constructions because conversants share a considerable amount of background information and, unlike writing, casual conversations are not distanced but immediate.

4.7 Conclusions

As Meyer (2012: 23) notes:

> In the pre-electronic era, textual analysis was largely a matter of analyzing 'static' texts: written texts existing only in printed form that had to be analyzed by hand.

He comments that the major earlier grammarians from this era, such as Otto Jespersen and Hendrik Poutsma, based their grammars on "primarily canonical written texts (e.g. novels)," from which they manually extracted relevant examples to illustrate the points of grammar that they discussed. Since this era, advances in corpus linguistics have resulted in the creation of many different types of corpora containing authentic texts ranging from spontaneous conversations to technical writing in the social and natural sciences. Because these texts are in an electronic format, they can be searched with software such as concordancing programs, and relevant linguistic constructions can be instantly retrieved. While it is certainly not the case that any particular linguistic item can be automatically retrieved instantly – many linguistic constructions are simply too complex for this type of "instant" retrieval – nevertheless, the process of corpus analysis has been greatly automated.

Concluding Remarks

The final chapter of the 1st edition of this book (hereafter ECL1), "Future Prospects in Corpus Linguistics," began with the following paragraph:

> In describing the complexity of creating a corpus, Leech (1998: xvii) remarks that "a great deal of spadework has to be done before the results [of a corpus analysis] can be harvested." Creating a corpus, he comments, "always takes twice as much time, and sometimes ten times as much effort" because of all the work that is involved in designing a corpus, collecting texts, and annotating them. And then, after a given period of time, Leech (1998: xviii) continues, the corpus becomes "out of date," requiring the corpus creator "to discard the concept of a static corpus of given length, and to continue to collect and store corpus data indefinitely into the future. . ." The process of analyzing a corpus may be easier than the description Leech (1998) gives above of creating a corpus, but still, many analyses have to be done manually, simply because we do not have the technology that can extract complex linguistic structures from corpora, no matter how extensively they are annotated. The challenge in corpus linguistics, then, is to make it easier both to create and analyze a corpus. What is the likelihood that this will happen?

In the context of current work done in corpus linguistics, the points made in the previous paragraph raise some interesting issues:

(1) *Planning a Corpus*: In ECL1, it is noted that "as more and more corpora have been created, we have gained considerable knowledge of how to construct a corpus that is balanced and representative and that will yield reliable grammatical information" (138). It is certainly the case that many recently created corpora are balanced and representative. Even some mega corpora, such as the one billion-word Corpus of Contemporary American English (COCA), contain various registers, such as fiction, speech, press reportage, and academic writing. But there are many challenges involved in creating "small and beautiful corpora," such as the British National Corpus (BNC) and the International Corpus of English (ICE). The situation is certainly understandable. It requires considerable resources to create balanced corpora, such as the BNC, whereas COCA

contains texts downloaded from websites, requiring much less effort than building a corpus such as the BNC.

(2) *Collecting and Computerizing Data*: In ECL1, it is noted that "because so many written texts are now available in computerized formats in easily accessible media. . . The collection and computerization of written texts has become much easier than in the past" (139). A similar situation exists in the present, and because the Web has grown even larger in recent years, plenty of texts can be downloaded. The situation with spoken texts was quite different: they had to be transcribed manually, a very time-consuming endeavor. In the present, there are transcriptions of different types of public speech, such as press conferences or talk shows, that can be downloaded and that are reasonably accurate. In fact, COCA contains quite a bit of public speech.

(3) *Annotating a Corpus*: Corpus annotation has changed considerably since ECL1. For instance, the opening of the section in ECL1 on "Structural Markup" makes reference to ASCII (American Standard Code for Information Interchange) text formats and standard generalized markup language (SGML). ASCII has since been replaced by Unicode, which has far more characters.

(4) *Tagging and Parsing a Corpus*: Lexically tagging a corpus has become quite routine, particularly because the accuracy of current tagging programs is quite high. Parsing a corpus is a more complicated undertaking, since larger structures, such as phrases and clauses, must be identified. Consequently, more post-editing has to be done after a corpus has been parsed.

(5) *Analyzing a Corpus*: Great strides have been made in developing software that can enhance the analysis of corpora. For instance, there are concordancing programs that can be used to identify and retrieve various grammatical constructions in a corpus, such as particular words or phrases. And if a corpus has been lexically tagged, concordancing programs can retrieve various grammatical structures: lexical items, such as nouns or verbs; phrases, such as noun phrases or verb phrases; and clauses, such as relative clauses or adverbial clauses.

In the future, we can expect further advances to enhance the creation and analysis of linguistic corpora. In particular, we can expect an increase in accuracy for certain corpus tools, such as improvement in the accuracy of software used to parse and lexically tag corpora, and the increased accuracy of tools, such as concordancing programs, used to retrieve constructions from corpora.

Discussion Topics

Below are a series of questions that explore the topics discussed in the first four chapters of the book. The questions go beyond fill in the blank- or short essay-types of responses and require respondents to think more deeply and critically in their answers. Consequently, there are no correct or incorrect answers.

Chapter 1

1. In the opening section of the chapter, a distinction is made between the "armchair linguist," whose sources of data are essentially sentences that he/she makes up, and the "corpus linguist," who believes that it is better to draw upon authentic data as the basis of linguistic analyses. In the book, it is noted that these two approaches are not necessarily mutually exclusive: the generative linguist, for instance, could very well work with examples that are corpus-based rather than examples that are made up. But is this true? Couldn't the generative linguist very well work with only made-up sentences, especially since his/her goal is to theorize about language, and spending additional time finding relevant data in a corpus is simply unnecessary?
2. What is the difference between "corpus" and "experimental" data? In the chapter, it is stated that these two different kinds of data "complement" each other. However, couldn't it also be argued that the two types of data are so different that they are incompatible? Drawing upon evidence in Section 1.3 of the chapter, argue that the two types of data are either compatible or incompatible.
3. Section 1.4 contains a discussion of how the corpus methodology has many different applications. For instance, it can be used to study language variation or to help in the creation of dictionaries. Select one of the areas described in this section and briefly discuss

how corpus methodology has changed the way that research is done in the area. You may want to focus your response on areas that are discussed in greater detail in Section 1.4 than some of the other areas.

4. Section 1.6 contains a discussion of how the theory of construction grammar can shed light on the "problematic" nature of apposition in English. Do you agree with this claim? What evidence exists to support the claim?

Chapter 2

1. The chapter opens with a discussion of three different types of corpora: the British National Corpus (BNC), the Corpus of Contemporary American English (COCA), and the Corpus of Early English Correspondence (CEEC). Below are some questions to consider:
   a. The BNC was released in 1994. Is it too out of date to be of use anymore? If not, what value might it have? For instance, how could it be compared with the BNC2014, a more modern corpus that is directly modeled after the BNC? Think of additional corpora it could be compared to.
   b. The COCA is a web-based corpus. It can be analyzed online. What advantages does it have being online rather than locally on a personal computer?
   c. The CEEC is actually a group of corpora containing letters written by individuals during various time periods, ranging from 1410 to 1663. The corpora were designed so that a range of sociolinguistic variables, such as age, social class, and gender, could be studied. Because these are historical corpora, what kinds of sociolinguistic variables would be especially worth studying? One point that is made in the chapters is that fewer letters written by women are in the corpus than letters written by men.
2. The idea of the Web as a corpus was initially a controversial notion, largely because at the time, more traditional corpora were the norm, and analyzing the Web, with its seemingly endless size and unknown content, contradicted the whole idea of what a corpus was thought to be. In the chapter, Biber, Egbert, and Davies (2015) describe work they did identifying the various content that texts on the web contained, ultimately concluding

that 29.2 percent of the texts they examined were hybrid in nature, having "multiple communicative purposes." Does it make sense to call the Web a corpus, a corpus that contains such a heterogeneous group of texts?

3. Technological advances as well as the development of the Web has made it easier to create much larger corpora than first-generation corpora, such as the Brown Corpus, which was one million words in length. How important is the actual size of a corpus? Does a "small and beautiful" corpus such as the Brown Corpus (ignoring its age) have any advantages over COCA, which is currently one billion words in length?
4. Let's say you wanted to create a corpus of newspaper editorials. How would you construct the corpus? For instance, would you want the corpus to contain a specific section of a newspaper (e.g. editorials). Or would you want more broad representation? Would you want complete texts, or only text excerpts? how would you ensure that the corpus is balanced? Consider other variables discussed in the chapter in your response.

Chapter 3

1. As is noted in the chapter, collecting and transcribing spoken language is one of the more labor-intensive parts of creating a corpus. But including spoken language in a corpus is very important because spoken language is the essence of human language, particularly spontaneous conversation. There are ways to get transcriptions of spontaneous speech, but mainly from conversations broadcast on radio or television, a highly restricted type of language. If you were considering creating a corpus of spontaneous conversations, how would you go about recording and transcribing them? Draw upon information in the chapter to write your response.
2. Let's say you want to create a corpus of newspaper editorials. How would you create a balanced corpus of them? For instance, how would you achieve gender balance? Would it be necessary to control for the types of newspapers from which you collect texts (e.g. broadsheets vs. tabloids)? What about other variables such as age and ethnicity? What are some variables whereby it would be too difficult to obtain information?
3. Go to the following website, which contains a much fuller description of the International Corpus of English (ICE) than the chapter

does: www.ice-corpora.uzh.ch/en/design.html. Would you consider the ICE corpora balanced corpora? Why or why not?

4. What is the difference between metadata and textual markup? Why is it important that a standard system, such as the Text Encoding Initiative, is developed to ensure that everyone follows a standardized system of annotation for representing metadata and textual annotation?

Chapter 4

1. Section 4.5 discusses how corpus analyses can be quantitative, qualitative, or a combination of the two (sometimes termed mixed methods). What's the difference between the three types of analysis? Is any one type of analysis better than another?
2. Conducting a corpus analysis is a multi-stage process, involving framing a research question, finding a suitable corpus or corpora for the analysis, extracting relevant information from the corpus or corpora chosen for the analysis, and integrating information from relevant articles or books into the analysis. Why are all these stages necessary? What would happen if, for instance, you did your analyses but did not integrate relevant research into the write-up (book, article, presentation) of your results?
3. Go to www.english-corpora.org/ and select a corpus from the list of corpora on the page. After you have selected a corpus, you will need to create an account to use the corpus. Once you have created an account, you should replicate some of the searches in Section 4.3 of the chapter but try your own searches too. You could also try BNCweb (http://corpora.lancs.ac.uk/BNCweb/) to search the British National Corpus online.
4. Section 4.6 describes Douglas Biber's use of multi-dimensional analysis as a way of finding linguistic differences between, for instance, speech and writing. Table 4.1 lists select words that predominated in two texts included in the section – one written and the other spoken. While words such as the pronoun *you* occurred overwhelmingly in the spoken text, *you* also occurred (though much less frequently) in the written text. What would motivate someone to use *I* in writing? Likewise, why might the nominalization *assessment*, which is much more frequent in writing, be used in conversation?

Appendix: Corpora

A

ARCHER Corpus: consists of various genres of British and American English covering the periods 1600–1999. (www.alc.manchester.ac.uk/linguistics-and-english-language/research/projects/archer/)

B

Bank of English Corpus: a monitor corpus initially used for linguistic analysis but that has been integrated into Word Banks online, a large corpus used by Collins Publishers for lexicographical work. (www.lancaster.ac.uk/fass/projects/corpus/cbls/corpora.asp#_Toc92298877)

Bergen Corpus of London Teenager English (COLT): contains the conversations of adolescents aged 13–17 from various social classes who live in different boroughs of London. (http://korpus.uib.no/icame/colt/)

BLOB-1931 Corpus: modeled after the London-Oslo- Bergen Corpus but with written texts covering the years 1928–1934. (https://varieng.helsinki.fi/CoRD/corpora/DOEC Dictionary of Old English Corpus)

British National Corpus: a 100 million-word corpus containing various spoken and written registers of British English dating back to the latter part of the twentieth century. (www.natcorp.ox.ac.uk/)

Brown Corpus (released in 1964): one million words in length of differing registers of written American English. (http://clu.uni.no/icame/manuals/)

BUMB (the Brown UMass Boston Corpus): contains collections of texts taken from American and British newspapers published in the 1930s that are comparable to the Brown Corpus. An internal corpus that has never been publicly released.

BYU Corpora: a set of many different corpora taken from various online resources (www.english-corpora.org/): see also entry under C for the Corpus of Contemporary American English.

C

CallHome Corpus: 120 spontaneous 30-minute phone conversations between intimates, either friends or members of the same family. (https://catalog.ldc.upenn.edu/LDC97S42)

CANCODE Corpus: a 5 million-word corpus containing various kinds of spoken English collected from differing locations in England. (www.nottingham.ac.uk/research/groups/cral/projects/cancode.aspx)

Child Language Data Exchange System, or CHILDES Corpus: includes transcriptions of children engaging in spontaneous conversations in English and other languages. (http://childes.talkbank.org)

Collins Corpus: a 4.5 billion-word monitor corpus used as the basis for creating the COBUILD dictionaries. (https://collins.co.uk/page/The+Collins+Corpus).

Corpus of Age and Gender (see Murphy 2010): an internal corpus not publicly available. (https://benjamins-com.ezproxy.lib.umb.edu/catalog/scl.38)

Corpus of Contemporary American English (COCA): a one billion-word corpus containing various registers of American English that can be searched online. (http://corpus.byu.edu/coca/)

Corpus of Early English Correspondence: consists of a collection of corpora containing various types of correspondence written in the fifteenth to seventeenth centuries. (www2.helsinki.fi/en/researchgroups/varieng/corpus-of-early-english-correspondence)

Corpus of Global Web-Based English (GloWbE): a 1.9 billion-word corpus containing samples of English from 20 different countries in which English is used. (http://corpus.byu.edu/glowbe/)

Corpus of Middle English Prose and Verse: contains the works of Chaucer and other Middle English writers. (www.hti.umich.edu/english/mideng/)

D

Diachronic Corpus of Present-Day Spoken English (DCPSE): 400,000 words of spoken English from ICE-GB and an

additional 400,000 spoken words from the London-Lund Corpus. (www.ucl.ac.uk/english-usage/projects/dcpse/index.htm)

Dictionary of Old English Corpus: a three million-word corpus containing all surviving Old English texts. (www.sheffield.ac.uk/library/cdfiles/doec#:~:text=The%20Dictionary%20of%20Old%20English,the%20collected%20works%20of%20Shakespeare)

E

Electronic *Beowulf*: an online version of *Beowulf*. (http://ebeowulf.uky.edu/)

English–Norwegian parallel Corpus: contains texts in both English and Norwegian translated into the other language, for instance English into Norwegian and Norwegian into English. (https://benjamins.com/catalog/scl.90)

Europarl Corpus (Release V7): transcriptions of 21 European languages taken from meetings of the European Parliament that were translated into English. (www.statmt.org/europarl/)

F

FLOB (The Freiburg LOB Corpus of British English) and FROWN (The Freiburg Brown Corpus of American English): corpora containing updated versions of the LOB (London-Oslo-Bergen Corpus) and Brown Corpus (which contain texts published in 1961) with comparable texts published in 1991. (www.lancaster.ac.uk/fss/courses/ling/corpus/blue/l02_1.htm)

FrameNet Project: created corpora containing various types of semantic tags, which mark features of what are called "semantic frames." (https://framenet.icsi.berkeley.edu/fndrupal/IntroPage)

H

Hansard Corpus of Canadian parliamentary proceedings: a parallel corpus containing debates translated into French and English. (https://spraakbanken.gu.se/lb/pedant/parabank/node6.html#:~:text=

Possibly%20the%20most%20well%2Dknown,official%20languages%2C%20English%20and%20French)

Helsinki Corpus: 1.5 million words representing Old English, Middle English, and Early Modern English. (https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/)

I

ICE-GB: the British component of the International Corpus of English: 1 million words of spoken and written British English that have been fully parsed and are searchable with a program called ICECUP 3.1. (www.ucl.ac.uk/english-usage/projects/ice-gb/)

ICLE Corpus (The International Corpus of Learner English): contains samples of written English produced by advanced learners of English as a foreign language from 25 different language backgrounds. (https://uclouvain.be/en/research-institutes/ilc/cecl/icle.html)

International Corpus of English: a collection of 1 million-word corpora from many national varieties of English (e.g. British English, Canadian English). Each component (with the exception of the American component, which has only the written component) contains the same types of spoken and written English (e.g. press reportage, scientific writing, spontaneous conversations). (www.ice-corpora.uzh.ch/en.html)

iWeb corpus: 14 billion words in length and containing texts taken from 22 million web pages. It is searchable online. (www.english-corpora.org/iweb/)

L

Lampeter Corpus of Early Modern English Tracts: a circa 1.1 million-word corpus consisting of complete texts ranging in length from 3,000 to 20,000 words. (http://korpus.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM)

LOB (London/Oslo-Bergen) Corpora (released in 1976): a one million-word corpus, modeled after the Brown Corpus, that contains various types of written British English published in 1961. (http://clu.uni.no/icame/manuals/)

London-Lund Corpus of Spoken English: one million words of various kinds of spoken British English covering the years

1959–1975. The samples included in the corpus were orthographically transcribed and annotated with prosodic information. (http://korpus.uib.no/icame/manuals/LONDLUND/INDEX.HTM)

London-Lund Corpus, LLC-2: a successor to the London-Lund Corpus (not yet publicly released). It contains a half-million words of conversations between speakers of British English and will enable comparisons with the original London-Lund Corpus. (https://projekt.ht.lu.se/llc2)

M

Map Task Corpus: 128 transcribed and annotated dialogues (http://groups.inf.ed.ac.uk/maptask/)

Michigan Corpus of Academic Spoken English (MICASE Corpus): consists of various types of spoken English used in academic contexts (e.g. advising sessions, colloquia) at the University of Michigan. (https://quod.lib.umich.edu/cgi/c/corpus/corpus?page=home;c=micase;cc=micase)

Microsoft Paraphrase Corpus: a corpus of sentence pairs containing human judgements as to whether the two pairs – one machine-generated and the other a paraphrase – are equivalent in meaning. (www.microsoft.com/en-us/download/details.aspx?id=52398)

N

Northern Ireland Transcribed Corpus of Speech: unscripted conversations of speakers of Hiberno-English. (https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/2475?show=full)

O

Old Bailey Corpus (1674–1913): composed of letters representing "speech-related documents" of court proceedings of the Old Bailey. (https://fedora.clarin-d.uni-saarland.de/oldbailey/)

P

Penn-Helsinki Parsed Corpus of Middle English and the Penn-Helsinki Parsed Corpus of Early Modern English: contain parsed versions of parts of the Helsinki Corpus. (www.ling.upenn.edu/hist-corpora/)

Penn Treebank: consists of various sentences that have been lexically tagged or skeletally parsed with one part of the Treebank markup of texts containing speech disfluencies. (https://link.springer.com/chapter/10.1007/978-94-010-0201-1_1)

Polytechnic of Wales Corpus: transcriptions of conversations between children (6–12 years of age) and an adult. Topics of discussion were the childrens' "favourite games or TV programmes." (http://shachi.org/resources/797)

Q

Quirk (Survey) Corpus: spoken and written British English collected between 1955 and 1985 and housed in file drawers at the Survey of English Usage. (www.ucl.ac.uk/english-usage/about/history.htm)

S

Santa Barbara Corpus of Spoken American English contains samples of various types of transcribed spoken American English. It is approximately 249,000 words in length. (www.linguistics.ucsb.edu/research/santa-barbara-corpus)

Sketch Engine: a resource containing not just various corpora (in English and other languages) but tools that can be used to analyze them. (www.sketchengine.co.uk/)

Spoken BNC 2014: the successor to the original British National Corpus; currently in progress. (http://corpora.lancs.ac.uk/bnc2014/)

Switchboard Corpus: 2,400 phone conversations (totaling 260 hours of speech) taking place between two individuals. (https://catalog.ldc.upenn.edu/ldc97s62)

The Synchronic English Web Corpus: miscellaneous texts taken from the Web. (www.corpusfinder.ugent.be/synchronic-english-web-corpus)

T

Treebank-3: a 100 million-word corpus containing a heterogeneous collection of texts, including 1 million words of text taken from the 1989 *Wall Street Journal* as well as tagged and parsed versions of the Brown Corpus. (https://catalog.ldc.upenn.edu/LDC99T42)

Trump Twitter Archive: a collection of Trump tweets covering the years 2009 and 2021. (www.thetrumparchive.com/)

Y

York English Corpus: a 1.5 million-word corpus that has been subjected to extensive analysis and that has yielded valuable information on dialect patterns (both social and regional) particular to this region of England. (www.researchgate.net/figure/Sample-design-of-the-York-English-Corpus_tbl1_227609144)

Bibliography

Aarts, Bas (2001) Corpus Linguistics, Chomsky and Fuzzy Tree Fragments. In Christian Mair and Marianne Hundt (eds.) (2001) *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi. 5–13.

Aarts, Bas and Charles F. Meyer (eds.) (1995) *The Verb in Contemporary English*. Cambridge: Cambridge University Press.

Aarts, Jan, Hans van Halteren, and Nelleke Oostdijk (1996) The TOSCA Analysis System. In C. Koster and E. Oltmans (eds.) *Proceedings of the First AGFL Workshop*. Nijmegen: CSI. 181–91.

Acuña, Juan Carlos Fariña (1996) *The Puzzle of Apposition: On So-Called Appositive Structures in English*. Santiago de Compostela: Universidade de Santiago de Compostela.

(2006) *A Constructional Network in Appositive Space*. Santiago de Compostela: Universidade de Santiago de Compostela.

Adolphs, Svenja and Ronald Carter (2013) *Spoken Corpus Linguistics: From Monomodal to Multimodal*. New York: Routledge.

Angouri, Jo (2018) Quantitative, Qualitative, Mixed or Holistic Research? Combining Methods in Linguistic Research. In Lia Litosseliti (ed.) *Research Methods in Linguistics*. London: Bloomsbury Publishing. 35–56.

Archer, Dawn, Andrew Wilson, and Paul Rayson (2002) Introduction to the USAS Category System. Benedict Project Report.

Aston, Guy and Lou Burnhard (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Bednarek, Monika (2014) 'Who Are You and Why Are You Following Us?' Wh-questions and Communicative Context in Television Dialogue. In John Flowerdew (ed.) *Discourse in Context (Contemporary Applied Linguistics 3)*. London/New York: Bloomsbury [formerly Continuum]. 49–70. www.monikabednarek.com/5.html

Bell, Alan (1988) The British Base and the American Connection in New Zealand Media English. *American Speech* 63: 326–44.

Biber, Douglas (1988) *Variation Across Speech and Writing*. New York: Cambridge University Press.

(1990) Methodological Issues Regarding Corpus-based Analyses of Linguistic Variation. *Literary and Linguistic Computing* 5: 257–69.

(1993) Representativeness in Corpus Design. *Literary and Linguistic Computing* 8: 241–57.

(1995) *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press.

(2019) Multi-Dimensional Analysis: A Historical Synopsis. In Tony Berber and Marcia Veriano (eds.) *Multi-Dimensional Analysis: Research Methods and Current Issues*. London: Bloomsbury Academic. 11–26.

Biber, Douglas and Jená Burges (2000) Historical Change in the Language Use of Women and Men: Gender Differences in Dramatic Dialogue. *Journal of English Linguistics* 28(1): 21–37.

Biber, Douglas and Bethany Gray (2016) *Grammatical Complexity in Academic English, Linguistic Change in Writing*. Cambridge: Cambridge University Press.

Biber, Douglas, Edward Finegan, and Dwight Atkinson (1994) ARCHER and Its Challenges: Compiling and Exploring a Representative Corpus of English Historical Registers. In Fries, Tottie, and Schneider (1993). 1–13.

Biber, Douglas, Susan Conrad, and Randi Reppen (1998) *Corpus Linguistics: Investigating Language Structure and Language Use*. Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan (1999) *The Longman Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas, Jesse Egbert, and Mark Davies (2015) Exploring the Composition of the Searchable Web: A Corpus-based Taxonomy of Web Registers. *Corpora* 10(1): 11–45.

Blachman, Edward, Charles F. Meyer, and Robert A. Morris (1996) The UMB Intelligent ICE Markup Assistant. In Greenbaum (1996a). 54–64.

Brezina, Vaclav (2018) *Statistics in Corpus Linguistics*. Cambridge: Cambridge University Press.

Brown, Penelope and Stephen Levinson (1987) *Politeness: Some Universals in Language Usage*. Cambridge: Cambridge University Press.

Chafe, Wallace (1994) *Discourse, Consciousness, and Time*. Chicago: University of Chicago Press.

(1995) Adequacy, User-Friendliness, and Practicality in Transcribing. In Leech, Myers, and Thomas (eds.) (1995). 54–61.

Chafe, Wallace, John Du Bois, and Sandra Thompson (1991) Towards a New Corpus of American English. In Karin Aijmer and Bengt Altenberg (eds.) *English Corpus Linguistics*. London: Longman. 64–91.

Chapelle, Carol A. (ed.) (2020) *The Concise Encyclopedia of Applied Linguistics*. Hoboken, NJ: John Wiley & Sons.

Chen, Xueliang, Yuanle Yan, and Jie Hu (2019) A Corpus-Based Study of Hillary Clinton's and Donald Trump's Linguistic Styles. *International*

Journal of English Linguistics 9(3): 13–22. www.ccsenet.org/journal/index.php/ijel/article/view/0/39048

Chomsky, Noam (1995) *The Minimalist Program*. Cambridge, MA: The MIT Press.

Clarke, Isobelle and Jack Grieve (2019) Stylistic Variation on the Donald Trump Twitter Account: A Linguistic Analysis of Tweets Posted Between 2009 and 2018. *PLoS ONE* 14(9): e0222062. doi:10.1371/journal.pone.0222062

Collins, Peter (1991a) *Cleft and Pseudo-Cleft Constructions in English*. Andover: Routledge.

(1991b) The Modals of Obligation and Necessity in Australian English. In Karin Aijmer and Bengt Altenberg (eds.) *English Corpus Linguistics*. London: Longman. 145–65.

Cook, Guy (1995) Theoretical Issues: Transcribing the Untranscribable. In Leech, Myers, and Thomas. 35–53.

Copestake, Ann (2016) Computational Linguistics. In Keith Allan (ed.) *The Routledge Handbook of Linguistics*. London and New York: Routledge. 485–501.

Corpus Encoding Standard (2000) www.cs.vassar.edu/CES/122.

Croft, William and D. Alan Cruse (2004) *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Crowdy, Steve (1993) Spoken Corpus Design. *Literary and Linguistic Computing* 8: 259–65.

Curzan, Anne (2003) *Gender Shifts in the History of English*. Cambridge: Cambridge University Press.

Davies, Alan (2003) *The Native Speaker: Myth and Reality*. Clevedon: Multilingual Matters.

Davies, Mark (2015) Corpora: An Introduction. In Douglas Biber and Randi Reppen (eds.) *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press. 11–31.

Dunning, Ted (1993) Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1): 61–74.

Ehlich, Konrad (1993) HIAT: A Transcription System for Discourse Data. In Jane Edwards and Martin Lampert (eds.) (1993) *Talking Data*. Hillside, NJ: Lawrence Erlbaum. 123–48.

Elsness, Johan (1982) That v. Zero Connective in English Nominal clauses. ICAME News: 1–45.

Elsness, J. (1997) *The Perfect and the Preterite in Contemporary and Earlier English*. Berlin & New York: Mouton de Gruyter.

Fang, Alex (1996) AUTASYS: Automatic Tagging and Cross-Tagset Mapping. In Greenbaum (ed.) (1996a). 110–24.

Fillmore, Charles (1992) Corpus Linguistics or Computer-Aided Armchair Linguistics. In Svartvik (ed.) (1992). 35–60.

Fillmore, Charles T., Paul Kay, and Mary Catherine O'Connor (1988) Regularity and Idiomaticity in Grammatical Constructions: The Case of Let Alone. *Language* 64(3): 501–83.

Francis, W. Nelson (1979) A Tagged Corpus—Problems and Prospects. In Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (eds.) *Studies in English Linguistics*. London: Longman. 192–209.

(1992) Language Corpora B.C. Jan Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin: Mouton de Gruyter. 17–31.

Francis, W. Nelson and Henry Kučera (1982) *Frequency Analysis of English Usage: Lexicon and Grammar*. Boston: Houghton Mifflin.

Fraser, Michael (1996) Tools and Techniques for Computer-assisted Biblical Studies. Paper delivered to the New Testament Research Seminar, Faculty of Theology, University of Oxford, June 1996. Available at: http://users.ox.ac.uk/~mikef/pubs/NT_Seminar_Oxford_Fraser_1996.html

Fries, Udo, Gunnel Tottie, and Peter Schneider (eds.) (1993) *Creating and Using English Language Corpora*. Amsterdam: Rodopi.

Garside, Roger, Geoffrey Leech, and Geoffrey Sampson (1987) *The Computational Analysis of English*. London: Longman.

Garside, Roger and Nicholas Smith (1997) A Hybrid Grammatical Tagger: CLAWS 4. In Garside, Leech and McEnery (1997). 102–21.

Garside, Roger, Geoffrey Leech, and Anthony McEnery (eds.) (1997) *Corpus Annotation*. London: Longman.

Garside, Roger, Geoffrey Leech, and Tamás Váradi (1992) *Lancaster Parsed Corpus*. Manual to accompany the Lancaster Parsed Corpus. http://khnt.hit.uib.no/icame/manuals/index.htm

Gatto, Maristella (2014) *Web as Corpus: Theory and Practice*. London: Bloomsbury.

A Gentle Introduction to SGML. In *Guidelines for Electronic Text Encoding and Interchange (TEI P3)* http://etext.lib.virginia.edu/bin/tei-tocs-p3?div=DIV1&id=SG

Gilquin, Gaëtanelle and Stefan T. Gries (2009) Corpora and Experimental Methods: A State-of-the-art Review. *Corpus Linguistics and Linguistic Theory* 5: 1–26.

Goldberg, Adele E. (2006) *Constructions at Work: The Nature of Generalization in Language*. New York: Oxford University Press.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier, and Magali Paquot (eds.) (2009) *International Corpus of Learner English, Version 2*. Louvain-la-Neuve, Belgium: Presses Universitaires de Louvain.

Gray, Bethany (2019) Tagging and Counting Linguistic Features for Multi-Dimensional Analysis. In T. Berber-Sardinha and M. Veirano (eds.) *Multi-dimensional Analysis: Research Methods and Current Issues*. Bloomsbury: Continuum. 43–66.

Greenbaum, Sidney (1984) Corpus Analysis and Elicitation Tests. In J. Aarts and W. Meijs (eds.) *Corpus Linguistics: Recent Developments in the Use of Computer Corpora*. Amsterdam: Rodopi. 195–201.

(ed.) (1996a) *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

(1996b) *The Oxford English Grammar*. Oxford: Oxford University Press.

Greenbaum, Sidney and Randolph Quirk (1970) *Elicitation Experiments in English: Linguistic Studies in Use and Attitude*. Coral Gables, FL: University of Miami Press.

Greenbaum, Sidney and Jan Svartvik (1990) The London-Lund Corpus of Spoken English. In Jan Svartvik (ed.) *The London-Lund Corpus of Spoken English: Description and Research*. Lund, Sweden: Lund University Press. 11–45.

Greene, Barbara B. and Gerald M. Rubin (1971) *Automatic Grammatical Tagging*. Technical Report. Department of Linguistics: Brown University.

Grice, H. Paul (1989) *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.

Gries, Stefan Th. (2017) *Quantitative Corpus Linguistics with R* (2nd edn.). Abingdon, Oxon and New York, NY: Routledge.

Gries, Stefan Th. and Andrea L. Berez (2017) Linguistic Annotation in/for Corpus Linguistics. In Nancy Ide and James Pustejovsky (eds.) *Handbook of Linguistic Annotation*. Dordrecht: Springer. 379–409.

Haegeman, Lilliane (1991) *Introduction to Government and Binding Theory*. Oxford: Blackwell.

Halliday, Michael and Ruqaiya Hasan (1976) Cohesion in *English*. London: Longman.

Hardie, Andrew (2014) Modest XML for Corpora: Not a Standard, But a Suggestion. *ICAME Journal* 38(1): 73–103. doi:10.2478/icame-2014-0004

Hasko, Victoria (2020) Qualitative Corpus Analysis. In Carol A. Chapelle (ed.) (2020) *The Concise Encyclopedia of Applied Linguistics*. Hoboken, NJ: John Wiley & Sons. 951–7.

Heenan, Charles H. (2002) Manual and Technology-Based Approaches to Using Classification for the Facilitation of Access to Unstructured Text. *(Unpublished Manuscript), Engineering Informatics Group, Stanford University.*

Hockey, Susan M. (2000) *Electronic Texts in the Humanities*. Oxford: Oxford University Press.

Hoffman, Sebastian (2007) From Web-Page to Mega-Corpus: The CNN Transcripts. In Marianne Hundt, Nadja Nesselhauf, and Carolin Biewer (eds.) *Corpus Linguistics and the Web*. Amsterdam: Rodopi. 69–85.

Hundt, Marianne and Geoffrey Leech (2012) "Small is Beautiful": On the Value of Standard Reference Corpora for Observing Recent

Grammatical Change. In Terttu Nevalainen and Elizabeth Closs Traugott (eds.) *The Oxford Handbook of the History of English*. Oxford: Oxford University Press. 175–88.

Jatav, Visaal, Teja, Ravi, Bharadwaj, Sudi, and Srinivasan, Mahesh (2017) Improving Part-of-Speech Tagging for NLP Pipelines. *CoRR, abs/1708.00241*.

Jespersen, Otto (1909–49) *A Modern English Grammar on Historical Principles*. Copenhagen: Munksgaard.

Kalton, Graham (1983) *Introduction to Survey Sampling*. Beverly Hills, CA: Sage.

Keay, Julia (2005) *Alexander the Corrector: The Tormented Genius whose Cruden's Concordence Unwrote the Bible*. Woodstock and New York: The Overlook Press.

Kennedy, Graeme (1998) *An Introduction to Corpus Linguistics*. London: Routledge.

(1996) Over Once Lightly. In Carol Percy, Charles F. Meyer, and Ian Lancashire (eds.) *Synchronic Corpus Linguistics*. Amsterdam: Rodopi. 253–62.

Kepser, Stephan and Marga Reis (eds.) (2005) *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*. Berlin: Mouton de Greyter.

Kilgarriff, Adam, Sue Atkins, and Michael Rundell (2007) BNC Design Model Past its Sell-by. In Proc. Corpus Linguistics. Birmingham, UK. Retrieved from www.kilgarriff.co.uk/Publications/2007-KilgAtkinsRundell-CL-Sellby.pdf.

Kilgarriff, Adam and Iztok Kosem (2012) Corpus Tools for Lexicographers. In Sylviane Granger and Magali Paquot (eds.) *Electronic Lexicography*. Oxford: Oxford University Press. 31–55.

Kirk, John (1992) The Northern Ireland Transcribed Corpus of Speech. In Gerhard Leitner (ed.) *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter. 65–73.

Kretzschmar, William, Charles F. Meyer, and Dominique Ingegneri (1997) Uses of Inferential Statistics in Corpus Linguistics. In Magnus Ljung (ed.) *Corpus Based Studies in English*. Amsterdam: Rodopi. 167–77.

Kučera, Henry (2002) Obituary for W. Nelson Francis. *Journal of English Linguistics*, 30(4), 306–9. doi:10.1177/007542402237878

Kučera, Henry and W. Nelson Francis (1967) *Computational Analysis of Present-Day American English*. Providence: Brown University Press.

Kytö, Merja (1991) *Variation and Diachrony, with Early American English in Focus. Studies on "Can"/"May" and "Shall"/"Will"*. University of Bamberg Studies in English Linguistics 28. Frankfurt am Main: Peter Lang.

(1996) *Manual to the Diachronic Part of the Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts* (3rd ed.). Department of English, University of Helsinki.

Labov, William (1972) The Transformation of Experience in Narrative Syntax. In *Language in the Inner City*. Philadelphia: University of Pennsylvania Press. 354–96.

Lakoff, George (1987) *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. Chicago: University of Chicago Press.

Landau, Sidney (1984) *Dictionaries: The Art and Craft of Lexicography*. New York: Charles Scribner.

Langacker, Ronald W. (2008) *Cognitive Grammar: A Basic Introduction*. Cambridge: Cambridge University Press.

Leech, Geoffrey (1983) *Principles of Pragmatics*. London: Longman.

(1992) Corpora and Theories of Linguistic Performance. In Jan Svartvik (ed.) *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin: Mouton de Gruyter. 105–22.

(1997) Grammatical Tagging. In Garside, Leech and McEnery (eds.) (1997). 19–33.

(1998) Preface. In Sylvianne Granger *Learner English on Computer*. London: Longman. xiv–xx.

Leech, Geoffrey, Roger Garside, and Eric Atwell (1983) The Automatic Grammatical Tagging of the LOB Corpus. *ICAME Journal* 7: 13–33.

Leech, Geoffrey, Greg Myers, and Jenny Thomas (eds.) (1995) *Spoken English on Computer*. Harlow, Essex: Longman.

Leech, Geoffrey and Elizabeth Eyes (1997) Syntactic Annotation: Treebanks. In Garside, Leech, and McEnery (eds.). 34–52.

Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith (2009) *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Leon, Fernando Sanchez and Amalio F. Nieto Serrano (1997) Retargeting a Tagger. In Garside, Leech and McEnery (eds.) (1997). 151–65.

Ljung, Magnus (ed.) (1997) *Corpus-based Studies in English*. Amsterdam: Rodopi.

Love, Robbie (2020) *Overcoming Challenges in Corpus Construction, The Spoken British National Corpus 2014*. New York and London: Routledge.

Lu, Xiaofei (2014) *Computational Methods for Corpus Annotation and Analysis*. New York: Springer.

Manning, Christopher D. (2002) Probabilistic Syntax. In Rens Bod, Jennifer Hay, and Stefani Jannedy (eds.) *Probabilistic Linguistics*. Cambridge, MA: The MIT Press. 289–342.

(2011) Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics? In A. F. Gelbukh *Computational Linguistics and Intelligent Text Processing*. CICLing. Lecture Notes in Computer Science, vol. 6608. Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-19400-9_14

Mair, Christian (1995) Changing Patterns of Complementation, and Concomitant Grammaticalisation, of the Verb Help in Present-Day British English. In Aarts and Meyer (eds.) (1995). 258–72.

Maniez, François (2000) Corpus of English proverbs and set phrases. Message posted on the "Corpora" List, January 24.

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz (1993) Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics* 19: 314–30.

Markus, Manfred (1997) Normalization of Middle English Prose in Practice. In Magnus Ljung *Corpus-Based Studies in English*. Amsterdam: Rodopi. 211–26.

McCarthy, Michael and Anne O'Keeffe (2010) Historical Perspective: What Are Corpora and How Have They Evolved? In Anne O'Keefe and Michael McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*. New York: Routledge. 3–13.

McEnery, Tony and Andrew Hardie (2012) *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

(2013) The History of Corpus Linguistics. In Keith Allan (ed.) *The Oxford Handbook of the History of Linguistics*. Oxford Handbooks Online: Oxford University Press. 727–45.

Meyer, Charles F. (1987) Apposition in English. *Journal of English Linguistics* 20(1): 101–21.

(1992) *Apposition in Contemporary English*. Cambridge: Cambridge University Press.

(2002) Pseudo-Titles in the Press Genre of Various Components of the International Corpus of English. In Randi Reppen, Susan M. Fitzmaurice, and Douglas Biber (eds.) *Using Corpora to Explore Linguistic Variation*. Vol. 9. Amsterdam: John Benjamins. 147–66.

(2009) The Empirical Tradition in Linguistics *Journal of English Linguistics* 37: 208–13.

(2009) Pre-Electronic Corpora. In Anke Lüdeling and Merja Kytö (eds.) *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter. 1–14.

(2012) Textual Analysis: From Philology to Corpus Linguistics. In Merja Kytö (ed.) *English Corpus Linguistics: Crossing Paths*. Amsterdam: Rodopi. 23–42.

(2014) A Diachronic Study of Pseudo-titles and Related Appositives in the Press Reportage of British and American Newspapers. In Eugene Green and Charles F. Meyer (eds.) *The Variability of Current World Englishes*. Berlin: Mouton de Gruyter. 239–56.

Mollin, Sandra (2007) The Hansard Hazard: Gauging the Accuracy of British Parliamentary Transcripts. *Corpora* 2(2): 187–210. doi:10.3366/cor.2007.2.2.187

Mönnink, Inga (1997) Using Corpus and Experimental Data: A Multi-Method Approach. In Ljung (ed.) (1997). 227–44.

Murphy, Bróna (2010) *Corpus and Sociolinguistics: Investigating Age and Gender in Female Talk*. Amsterdam: John Benjamins.

Nevalainen, Terttu (2000) Gender Differences in the Evolution of Standard English: Evidence from the Corpus of Early English Correspondence. *Journal of English Linguistics* 28(1): 38–59.

Nevalainen, Terttu and Helena Raumolin-Brunberg (eds.) (1996) *Sociolinguistics and Language History: Studies Based on the Corpus of Early English Correspondence*. Amsterdam: Rodopi.

New York Times (2009, Sept. 11) "The Cookbook Wars: Judge Rejects Copyright Suit Against Jessica Seinfeld." Retrieved from https://artsbeat.blogs.nytimes.com/2009/09/11/the-cookbook-wars-judge-rejects-copyright-suit-against-jessica-seinfeld/

Oakes, Michael P. (1998) *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi, Vincent (1998) *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Oostdijk, Nelleke (1991) *Corpus Linguistics and the Automatic Analysis of English*. Amsterdam and Atlanta, GA: Rodopi.

Pahta, Päivi and Saara Nevanlinna (1997) Re-phrasing in Early English. Expository Apposition with an Explicit Marker from 1350 to 1710. In Matti Rissanen, Merja Kytö, and Kirsi Heikkonen (eds.) *English in Transition: Corpus-based Studies in English Linguistics and Genre Styles. Topics in English Linguistics* 23. Berlin & New York: Mouton de Gruyter. 121–83.

Pepper, Steven (2000) The Whirlwind Guide to SGML & XML Tools and Vendors. www.infotek.no/sgmltool/guide.htm

Pollard, Carl and Ivan A. Sag (1994) *Head-Driven Phrase Structure Grammar*. Chicago, IL: University of Chicago Press.

Porter, Nick and Akiva Quinn (1996) Developing the ICE Corpus Utility Program. In Greenbaum (ed.) (1996a). 79–91.

Powell, Christina and Rita C. Simpson (2001) Collaboration Between Corpus Linguists and Digital Librarians for the MICASE Web Search Interface. In R. Simpson and J. Swales (eds.) *Corpus Linguistics in North America*. Ann Arbor, MI: University of Michigan. 32–47.

Prescott, Andrew (1997) The Electronic Beowulf and Digital Restoration. *Literary and Linguistic Computing* 12: 185–95.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Renouf, Antoinette (1987) Corpus Development. In Sinclair (ed.) (1987). 1–40.

Rissanen, Matti (1992) The Diachronic Corpus as a Window to the History of English. In J. Svartvik (ed.) *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter. 185–205.

(2000) The World of English Historical Corpora: From Cædmon to the Computer Age. *Journal of English Linguistics* 28(1): 7–20.

Robinson, Peter (1998) New Methods of Editing, Exploring, and Reading *The Canterbury Tales*. http://cts.dmu.ac.uk/repository/robinson-1998/index.html

Ruskin, John (1866) *The Crown of Wild Olive*. London: Smith, Elder & Co.

Samuelsson, Christer and Voutilainen, Atro (1997) *Comparing a Linguistic and a Stochastic Tagger*. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics: Madrid, Spain. 246–53.

Schmied, Josef (1996) Second-Language Corpora. In Greenbaum (ed.) (1996a). 182–96.

Schmied, Josef and Claudia Claridge (1997) Classifying Text- or Genre-Variation in the Lampeter Corpus of Early Modern English Texts. In Raymond Hickey, Merja Kytö, Ian Lancashire, and Matti Rissanen (eds.) *Tracing the Trail of Time: Proceedings of the Diachronic Corpora Workshop, Toronto (Canada) May 1995*. Amsterdam, Atlanta: Rodopi. 119–35.

Shastri, S. V. (1988) The Kolhapur Corpus of Indian English and Work Done on its Basis So Far. *ICAME Journal* 2: 15–26.

Sigley, Robert J. (1997) "Choosing Your Relatives: Relative Clauses in New Zealand English." Unpublished PhD thesis. Wellington: Department of Linguistics, Victoria University of Wellington.

Simpson, Rita, Bret Lucka, and Janine Ovens (2000) Methodological Challenges of Planning a Spoken Corpus with Pedagogical Outcomes. In *Rethinking Language Pedagogy from a Corpus Perspective: Papers from the Third International Conference on Teaching and Language Corpora*. Frankfurt am Main: Peter Lang. 43–9.

Sinclair, John (ed.) (1987) *Looking Up: An Account of the COBUILD Project*. London: Collins.

Sinclair, John, Susan Jones, and Robert Daley (2004) *English Collocation Studies: The OSTI Report*. London: Bloomsbury Publishing.

Stenström, Anna-Brita and Gisle Andersen (1996) More Trends in Teenage Talk: A Corpus-Based Investigation of the Discourse Items cos and innit. In Carol Percy, Charles F. Meyer, and Ian Lancashire (eds.) *Synchronic Corpus Linguistics*. Amsterdam: Rodopi. 189–203.

Svartvik, Jan (ed.) (1990) *The London-Lund Corpus of Spoken English*. Lund, Sweden: Lund University Press

(1992) *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter.

Svartvik, Jan and Randolph Quirk (eds.) (1980) *A Corpus of English Conversation*. Lund, Sweden: Lund University Press.

Sweet, Henry (1891–98) *A New English Grammar*. Oxford: Clarendon Press.

Tagliamonte, Sali (1998) *Was/Were* Variation across the Generations: View from the City of York. *Language Variation and Change* 10(2): 153–91.

Tagliamonte, Sali and Helen Lawrence (2000) "I Used to Dance, but I don't Dance Now": The Habitual Past in English. *Journal of English Linguistics* 28(4): 324–53.

Tannen, Deborah (1989) *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*. Cambridge: Cambridge University Press.

Tapanainen, Pasi and Timo Järvinen (1997) A Non-Projective Dependency Parser. www.conexor.fi/anlp97/anlp97.html [also published in Procs. *ANLP-97*. ACL. Washington, DC]

Tatlock, John S. P. and Arthur Garfield Kennedy (1963) *A Concordance to the Complete Works of Geoffrey Chaucer and to the "Romaunt of the Rose." 1927*. Washington: The Carnegie Institute of Washington.

Thompson, Henry S. and David McKelvie (1996) A software architecture for simple, efficient SGML applications. Paper presented at SGML Europe '96 Munich. www.ltg.ed.ac.uk/software/nsl/sgml-europe/talk.html.

Thompson, Henry S., Anne H. Anderson, and Miles Bader (1995) Publishing a Spoken and Written Corpus on CD-ROM: The HCRC Map Task Experience. In Leech, Myers, and Thomas (eds.) (1995). 168–80.

Tognini-Bonelli, Elena (2001) *Corpus Linguistics at Work* Vol. 6. Philadelphia: J. Benjamins.

Voutilainen, Atro (1999) A Short History of Tagging. In Hans van Halteren (ed.) *Syntactic Word Class Tagging. Text, Speech and Language Technology* Vol. 9. Dordrecht: Springer. 9–21.

Wallis, Sean (2003) Completing Parsed Corpora: From Correction to Evolution. In Anne Abeillé (ed.) *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Springer. 61–71.

(2021) *Statistics in Corpus Linguistics Research*. New York and London: Routledge.

Xiong, W. et al. (2017) Toward Human Parity in Conversational Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25(12): 2410–23. doi:10.1109/TASLP.2017.2756440

Index