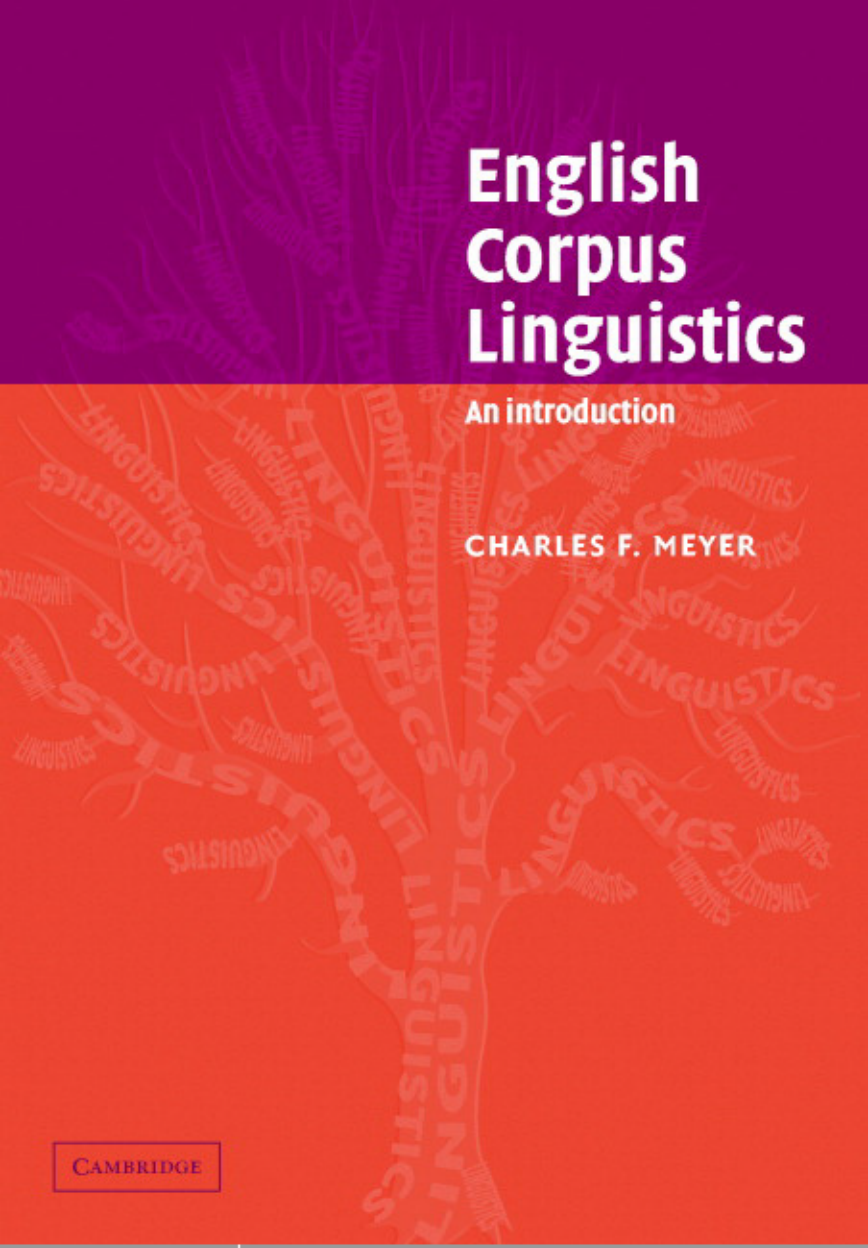
English Corpus Linguistics

An introduction

CHARLES F. MEYER



CAMBRIDGE

CAMBRIDGE

more information - www.cambridge.org/0521808790

English Corpus Linguistics
An Introduction

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

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

STUDIES IN ENGLISH LANGUAGE

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

Already published

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

Charles F. Meyer
Apposition in contemporary English

Jan Firbas
Functional sentence perpective in written and spoken communication

Izchak M. Schlesinger
Cognitive space and linguistic case

Katie Wales
Personal pronouns in present-day English

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

STUDIES IN ENGLISH LANGUAGE

English Corpus Linguistics
An Introduction

English Corpus Linguistics
An Introduction

CHARLES F. MEYER
University of Massachusetts at Boston

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

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

http://www.cambridge.org

First published in printed format 2002

ISBN 0-511-04200-0 eBook (netLibrary)
ISBN 0-521-80879-0 hardback
ISBN 0-521-00490-X paperback

To Libby and Freddie

Contents

Preface

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

1 Corpus analysis and linguistic theory

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

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

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

1.1 Linguistic theory and description

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

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

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

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

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

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

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

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

why these two groups of linguists have had such a difficult time communicating and valuing each other's work. As Fillmore (1992: 35) jokes, when the corpus linguist asks the theoretician (or "armchair linguist") "Why should I think that what you tell me is *true*?", the generative grammarian replies back "Why should I think that what you tell me is *interesting*?" (emphasis added). Of primary concern to the corpus linguist is an accurate description of language; of importance to the generative grammarian is a theoretical discussion of language that advances our knowledge of universal grammar.

Even though the corpus linguist places a high priority on descriptive adequacy, it is a mistake to assume that the analysis of corpora has nothing to offer to generative theory in particular or to theorizing about language in general. The main argument against the use of corpora in generative grammar, Leech (1992) observes, is that the information they yield is biased more towards performance than competence and is overly descriptive rather than theoretical. However, Leech (1992: 108) argues that this characterization is 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).

Despite Leech's claims, it is unlikely that corpora will ever be used very widely by generative grammarians, even though some generative discussions of language have been based on corpora and have demonstrated their potential for advancing generative theory. Working within the framework of government and binding theory (the theory of generative grammar preceding minimalist theory), Aarts (1992) used sections of the corpus housed at the Survey of English Usage at University College London to analyze "small clauses" in English, constructions like *her happy* in the sentence *I wanted her happy* that can be expanded into a clausal unit (*She is happy*). By using the London Corpus, Aarts (1992) was not only able to provide a complete description of small clauses in English but to resolve certain controversies regarding small clauses, such as establishing the fact that they are independent syntactic units rather than simply two phrases, the first functioning as direct object and the second as complement of the object.

Haegeman (1987) employed government and binding theory to analyze empty categories (i.e. positions in a clause where some element is missing) in a specific genre of English: recipe language. While Haegeman's investigation is not based on data from any currently available corpus, her analysis uses the type of data quite commonly found in corpora. Haegeman (1987) makes the very interesting claim that parametric variation (such as whether or not a language exhibits pro-drop) does not simply distinguish individual languages from one another but can be used to characterize regional, social, or register variation within a

particular language. She looks specifically at examples from the genre (or register) of recipe language that contain missing objects (marked by the letters [a], [b], etc. in the example below):

(1) Skin and bone chicken, and cut [a] into thin slices. Place [b] in bowl with mushrooms. Purée remaining ingredients in blender, and pour [c] over chicken and mushrooms. Combine [d] and chill [e] well before serving. (Haegeman 1987: 236–7)

Government and binding theory, Haegeman (1987: 238) observes, recognizes four types of empty categories, and after analyzing a variety of different examples of recipe language, Haegeman concludes that this genre contains one type of empty category, *wh*-traces, not found in the core grammar of English (i.e. in other genres or regional and social varieties of English).

What distinguishes Haegeman's (1987) study from most other work in generative grammar is that she demonstrates that theoretical insights into universal grammar can be obtained by investigating the periphery of a language as well as the core. And since many corpora contain samples of various genres within a language, they are very well suited to the type of analysis that Haegeman (1987) has conducted. Unfortunately, given the emphasis in generative grammar on investigations of the core of a language (especially as reflected in Chomsky's recent work in minimalism), corpora will probably never have much of a role in generative grammar. For this reason, corpora are much better suited to functional analyses of language: analyses that are focused not simply on providing a formal description of language but on describing the use of language as a communicative tool.

1.2 Corpora in functional descriptions of language

Even though there are numerous functional theories of language, all have a similar objective: to demonstrate how speakers and writers use language to achieve various communicative goals.[1]

Because functionalists are interested in language as a communicative tool, they approach the study of language from a markedly different perspective than the generative grammarian. As "formalists," generative grammarians are primarily interested in describing the form of linguistic constructions and using these descriptions to make more general claims about Universal Grammar. For instance, in describing the relationship between *I made mistakes*, a sentence in the active voice, and its passive equivalent, *Mistakes were made by me*, a generative grammarian would be interested not just in the structural changes in word order between actives and passives in English but in making more general claims about the movement of constituents in natural language. Consequently, the movement of noun phrases in English actives and passives is part

[1] Newmeyer (1998: 13–18) provides an overview of the approaches to language study that various functional theories of language take.

of a more general process termed "NP [noun phrase] – movement" (Haegeman 1991: 270–3). A functionalist, on the other hand, would be more interested in the communicative potential of actives and passives in English. And to study this potential, the functionalist would investigate the linguistic and social contexts favoring or disfavoring the use of, say, a passive rather than an active construction. A politician embroiled in a scandal, for instance, might choose to utter the agentless passive *Mistakes were made* rather than *I made mistakes* or *Mistakes were made by me* because the agentless passive allows the politician to admit that something went wrong but at the same time to evade responsibility for the wrong-doing by being quite imprecise about exactly who made the mistakes.

Because corpora consist of texts (or parts of texts), they enable linguists to contextualize their analyses of language; consequently, corpora are very well suited to more functionally based discussions of language. To illustrate how corpora can facilitate functional discussions of language, this section contains an extended discussion of a functional analysis of elliptical coordinations in English based on sections of the Brown Corpus and the American component of the International Corpus of English (ICE). The goal of the analysis (described in detail in Meyer 1995) was not simply to describe the form of elliptical coordinations in speech and writing but to explain why certain types of elliptical coordinations are more common than others, why elliptical coordinations occur less frequently in speech than in writing, and why certain types of elliptical coordinations are favored more in some written genres than others.

The study was based on a 96,000-word corpus containing equal proportions of different types of speech and writing: spontaneous dialogues, legal cross examinations, press reportage, belles lettres, learned prose, government documents, and fiction. These genres were chosen because they are known to be linguistically quite different and to have differing functional needs. Government documents, for instance, are highly impersonal. Consequently, they are likely to contain linguistic constructions (such as agentless passives) that are associated with impersonality. Spontaneous dialogues, on the other hand, are much more personal, and will therefore contain linguistic constructions (such as the personal pronouns *I* and *we*) advancing an entirely opposite communicative goal. By studying genres with differing functional needs, one can take a particular linguistic construction (such as an elliptical coordination), determine whether it has varying frequencies and uses in different genres, and then use this information to determine why such distributions exist and to isolate the function (or communicative potential) of the construction.

In an elliptical coordination, some element is left out that is recoverable within the clause in which the ellipsis occurs. In the sentence *I wrote the introduction and John the conclusion* the verb *wrote* is ellipted in the second clause under identity with the same verb in the first clause. There are various ways to describe the different types of ellipsis occurring in English and other languages. Sanders (1977) uses alphabetic characters to identify the six different positions in which

ellipsis can occur, ranging from the first position in the first clause (position A) to the last position in the second clause (position F):

A B C & D E F

Although there is disagreement about precisely which positions permit ellipsis in English, most would agree that English allows ellipsis in positions C, D, and E. Example (2) illustrates C-Ellipsis: ellipsis of a constituent at the end of the first clause (marked by brackets) that is identical to a constituent (placed in italics) at the end of the second clause.

(2) The author wrote [ ] and the copy-editor revised *the introduction to the book*.

Examples (3) and (4) illustrate D- and E-Ellipsis: ellipsis of, respectively, the first and second parts of the second clause.

(3) *The students* completed their course work and [ ] left for summer vacation.
(4) Sally *likes* fish, and her mother [ ] hamburgers.

The first step in studying the functional potential of elliptical coordinations in English was to obtain frequency counts of the three types of elliptical coordinations in the samples of the corpus and to explain the frequency distributions found. Of the three types of ellipsis in English, D-Ellipsis was the most frequent, accounting for 86 percent of the elliptical coordinations identified in the corpus. In contrast, both C- and E-Ellipsis were very rare, occurring in, respectively, only 2 percent and 5.5 percent of the elliptical coordinations.[2] These frequency distributions are identical to those found by Sanders (1977) in a survey he conducted of the frequency of ellipsis types in a variety of different languages. For instance, Sanders (1977) found that while all of the languages of the world allow D-Ellipsis, far fewer permit C-Ellipsis.

To explain typological distributions such as this, Sanders (1977) invokes two psycholinguistic constraints: the suspense effect (as Greenbaum and Meyer 1982 label it) and the serial position effect. Briefly, the suspense effect predicts that ellipsis will be relatively undesirable if the site of ellipsis precedes the antecedent of ellipsis, since the suspense created by the anticipation of the ellipted item places a processing burden on the hearer or reader. C-Ellipsis is therefore a relatively undesirable type of ellipsis because the antecedent of ellipsis (*the introduction to the book* in example 2) comes after the ellipsis in position C at the end of the first clause. D- and E-Ellipsis, on the other hand, are more desirable than C-Ellipsis because neither ellipsis type violates the suspense effect: for both types of ellipsis, the antecedent of ellipsis occurs in the first clause (position A for D-Ellipsis and position B for E-Ellipsis) in positions prior to ellipsis in the D- and E-positions in the second clause.

[2] The remaining 6.5 percent of elliptical coordinations consisted of constructions exhibiting more than one type of ellipsis and therefore no tendency towards any one type of ellipsis. For example, the example below contains both C- and D-Ellipsis: ellipsis of the direct object in the first clause and subject of the second clause.

(i) *We*$_1$ tried out [ ]$_2$ and [ ]$_1$ then decided to buy *the car*$_2$.

Table 1.1 *The favorability of C-, D-, and E- Ellipsis*

Ellipsis type	Suspense effect	Serial position effect
D-Ellipsis	F	F
E-Ellipsis	F	L
C-Ellipsis	L	L

F = favorable
L = less favorable

The serial position effect is based on research demonstrating that when given memory tests, subjects will remember items placed in certain positions in a series better than other positions. For instance, subjects will recall items placed first in a series more readily and accurately than items placed in the middle of a series. The results of serial learning experiments can be applied to the six positions in a coordinated construction (A–F) and make predictions about which antecedent positions will be most or least conducive to memory retention and thus favor or inhibit ellipsis. Position A, the antecedent position for D-Ellipsis (see example 3), is the position most favorable for memory retention. Consequently, D-Ellipsis will be the most desirable type of ellipsis according to the serial position effect. The next most favorable position for memory is position B, the antecedent position for E-Ellipsis, making this type of ellipsis slightly less desirable than D-Ellipsis. And increasingly less desirable for memory retention is the F-position, the antecedent position for C-Ellipsis, resulting in this type of ellipsis being the least desirable type of ellipsis in English.

Working together, the Suspense and Serial Position Effects make predictions about the desirability of ellipsis in English, predictions that match exactly the frequency distributions of elliptical coordinations found in the corpora. Table 1.1 lists the three types of ellipsis in English and the extent to which they favorably or unfavorably satisfy the suspense and serial position effects. D-Ellipsis quite favorably satisfies both the suspense and serial position effects, a fact offering an explanation of why D-Ellipsis was the most frequent type of ellipsis in the corpus. While E-Ellipsis satisfies the suspense effect, it less favorably satisfies the serial position effect, accounting for its less frequent occurrence in the corpus than D-Ellipsis. However, E-Ellipsis was more frequent than C-Ellipsis, a type of ellipsis that satisfies neither the suspense nor the serial position effect and was therefore the least frequent type of ellipsis in the corpus.

While the suspense and serial position effects make general predictions about the favorability or unfavorability of the three ellipsis types in English, they fail to explain the differing distributions of elliptical coordinations in speech and writing and in the various genres of the corpora. In speech, of the constructions in which ellipsis was possible, only 40 percent contained ellipsis, with the remaining 60 percent containing the full unreduced form. In writing, in contrast, ellipsis

was much more common: 73 percent of the constructions in which ellipsis was possible contained ellipsis, with only 27 percent containing the full unreduced form. To explain these frequency differences, it is necessary to investigate why repetition (rather than ellipsis) is more necessary in speech than in writing.

The role of repetition in speech is discussed extensively by Tannen (1989: 47–53), who offers a number of reasons why a construction such as (5) below (taken from a sample of speech in the American component of ICE) is more likely to occur in speech than in writing.

(5) Yeah so *we got* that and *we got* knockers and *we got* bratwurst and *we got* <unintelligible>wurst or kranzwurst or something I don't know. (ICE-USA-S1A-016)

In (5), there are four repetitions of a subject and verb (*we got*) in the D-position that could have been ellipted rather than repeated. But in this construction, repetition serves a number of useful purposes quite unique to speech. First, as Tannen (1989: 48) observes, the repetition allows the speaker to continue the flow of the discourse "in a more efficient, less energy-draining way" by enabling him/her to continue speaking without worrying about editing what is being said and getting rid of redundancies, a task that would greatly slow down the pace of speech. At the same time, repetition is beneficial to the hearer "by providing semantically less dense discourse" (p. 49), that is, discourse containing an abundance of old rather than new information. Moreover, repetition can create parallel structures (as it does in example 5), and as many researchers have noted, parallelism is a very common device for enhancing the cohesiveness of a discourse.

In addition to having a different distribution in speech and writing, elliptical coordinations also had different distributions in the various genres of writing that were investigated. If the genres of fiction and government documents are compared, very different patterns of ellipsis can be found. In fiction, D-Ellipsis constituted 98 percent of the instances of ellipsis that were found. In government documents, on the other hand, D-Ellipsis made up only 74 percent of the instances of ellipsis, with the remaining 26 percent of examples almost evenly divided between C-Ellipsis and E-Ellipsis.

The high incidence of D-Ellipsis in fiction can be explained by the fact that fiction is largely narration, and narrative action, as Labov (1972: 376) has shown, is largely carried forth in coordinate sentences. These sentences will often have as subjects the names of characters involved in the narrative action, and as these names are repeated, they will become candidates for D-Ellipsis. For instance, in example (6) below (which was taken from a sample of fiction in the Brown Corpus), the second sentence (containing two coordinated clauses) begins with reference to a male character (*He*) at the start of the first clause, a reference that is repeated at the start of the second clause, leading to D-Ellipsis rather than repetition of the subject. Likewise, the last two sentences (which also consist of coordinated clauses) begin with references to another character (*Virginia* initially and then *She*), which are repeated and ellipted in the D-positions of subsequent clauses.

(6) The days seemed short, perhaps because his routine was, each day, almost the same. *He* rose late and [ ] went down in his bathrobe and slippers to have breakfast either alone or with Rachel. *Virginia* treated him with attention and [ ] tried to tempt his appetite with special food: biscuits, cookies, candies – the result of devoted hours in the tiled kitchen. *She* would hover over him and, looking like her brother, [ ] anxiously watch the progress of Scotty's fork or spoon. (K01 610–80)

Although the government documents in the corpus contained numerous examples of D-Ellipsis, they contained many more examples of C-Ellipsis than the samples of fiction did. One reason that C-Ellipsis occurred more frequently in government documents is that this type of construction has a function well suited to government documents. As Biber (1988) has noted, the genre in which government documents can be found, official documents, has a strong emphasis on information, "almost no concern for interpersonal or affective content" (p. 131), and a tendency towards "highly explicit, text-internal reference" (p. 142).

Instances of C-Ellipsis quite effectively help government documents achieve these communicative goals. First of all, because government documents are so focused on content or meaning, they are able to tolerate the stylistic awkwardness of constructions containing C-Ellipsis. In example (7) below (taken from a government document in the Brown Corpus), there is a very pronounced intonation pattern created by the C-Ellipsis, resulting in pauses at the site of ellipsis and just prior to the ellipted construction that give the sentence a rather abrupt and awkward intonation pattern.

(7) Each applicant is required to own [ ] or have sufficient interest in *the property to be explored*. (H01 1980–90)

This awkwardness is tolerated in government documents because of the overriding concern in this genre for accuracy and explicitness. An alternative way to word (7) would be to not ellipt the noun phrase in the C-position but instead to pronominalize it at the end of the second clause:

(8) Each applicant is required to own *the property to be explored* or have sufficient interest in *it*.

However, even though this wording results in no confusion in this example, in general when a third-person pronoun is introduced into a discourse, there is the potential that its reference will be ambiguous. If, in the case of (7), ellipsis is used instead of pronominalization, there is no chance of ambiguity, since the constraints for ellipsis in English dictate that there be only one source for the ellipsis in this sentence (the noun phrase *the property to be explored* in the second clause). Consequently, through ellipsis rather than pronominalization, the communicative goal of explicitness in government documents is achieved.

The discussion of coordination ellipsis in this section provides further evidence that corpus-based analyses can achieve "explanatory adequacy": the results of the study establish a direct relationship between the frequency of the various types of elliptical coordinations across the languages of the world

and their overall frequency in English. More importantly, however, the analysis provides principled "functional" explanations for these frequency distributions in English: certain kinds of elliptical coordinations place processing burdens on the hearer/reader, thus making their overall frequency less common; at the same time, the less common constructions are sometimes necessary because they are communicatively necessary in certain contexts (e.g. the need in government documents to use a rare type of ellipsis, C-ellipsis, because this kind of construction prevents potential ambiguity that might occur with an alternative full-form containing a third-person pronoun).

Although not all corpus studies are as explicitly functional as the study of coordination ellipsis in this section, all corpus-based research is functional in the sense that it is grounded in the belief that linguistic analysis will benefit if it is based on real language used in real contexts. And as the next section will demonstrate, this methodological principle has influenced how research is conducted in numerous linguistic disciplines.

1.3 Corpus-based research in linguistics

Linguists of all persuasions have discovered that corpora can be very useful resources for pursuing various research agendas. For instance, many lexicographers 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 more systematic study of the evolution of English and that enable historical linguists to investigate issues that have currency in modern linguistics, such as the effects of gender on language usage 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. To illustrate how corpora have affected research in linguistics, the remainder of this chapter provides an overview of the various kinds of corpus-based research now being conducted in various linguistic disciplines.[3]

1.3.1 Grammatical studies of specific linguistic constructions

The study of coordination ellipsis in the previous section illustrated a very common use of corpora: to provide a detailed study of a particular grammatical construction that yields linguistic information on the construction,

[3] The following sections do not provide an exhaustive listing of the research conducted in the various areas of linguistics that are discussed. For a comprehensive survey of corpus-based research, see either Bengt Altenberg's online bibliography: –1989: http://www.hd.uib.no/icame/icame-bib2.txt; 1990–8: http://www.hd.uib.no/icame/icame-bib3.htm; or Michael Barlow's: http://www.ruf.rice.edu/~barlow/refn.html.

such as the various forms it has, its overall frequency, the particular contexts in which it occurs (e.g. speech rather than writing, fiction rather than spontaneous dialogues, and so forth), and its communicative potential.

Corpus-based research of this nature has focused on the use and structure of many different kinds of grammatical constructions, such as appositives in contemporary English (Meyer 1992) and earlier periods of the language (Pahta and Nevanlinna 1997); clefts and pseudo-clefts (Collins 1991b); infinitival complement clauses (Mair 1990); past and perfective verb forms in various periods of English (Elsness 1997); the modals *can/may* and *shall/will* in early American English (Kytö 1991); and negation (Tottie 1991) (see the ICAME Bibliography for additional studies).

To investigate the use and structure of a grammatical construction, most have found it more profitable to investigate constructions that occur relatively frequently, since if a construction occurs too infrequently, it is often hard to make strong generalizations about its form and usage. For instance, in the discussion of coordination ellipsis in the previous section, the infrequent occurrence of instances of E-Ellipsis (e.g. *Joe's a vegetarian, and Sally a carnivore*) helped make the theoretical point that if a particular grammatical construction occurs rarely in the world's languages, in those languages in which it does occur, it will have a very infrequent usage. At the same time, the lack of many examples of E-Ellipsis made it difficult to make strong generalizations about the usage of this construction in English. In many respects, this problem is a consequence of the relatively small corpus upon which the study of coordination ellipsis was based. For this reason, to study some linguistic constructions, it will often be necessary to study very large corpora: the British National Corpus, for instance (at 100 million words in length), rather than the Brown Corpus (at one million words in length). However, for those constructions that do occur frequently, even a relatively small corpus can yield reliable and valid information. To illustrate this point, it is instructive to compare two studies of modal verbs in English – Coates (1983) and Mindt (1995) – whose results are similar, even though the studies are based on very different sized corpora.

Coates (1983) was one of the earlier corpus studies of modals and was based on two corpora totaling 1,725,000 words: the Lancaster Corpus (a precursor to the LOB Corpus of written British English) and sections of the London Corpus containing speech, letters, and diaries. Coates (1983) used these two corpora to describe the different distributions of modals in writing and speech and the more frequent meanings associated with the individual modals. Mindt's (1995) study of modals was based on a much larger group of corpora that together totaled 80 million words of speech and writing: the Brown and LOB corpora, sections of the London–Lund Corpus containing surreptitiously recorded speech, the Longman–Lancaster Corpus, and CD-ROMS containing newspaper articles from *The Times* and the *Independent*. Mindt (1995) used these corpora not only to study the form and meaning of modals but to provide a comprehensive view

of the verb phrase in English based on the approximately 30,000 verb phrases he identified in his corpora.

Although the size of Coates' (1983) and Mindt's (1995) corpora is drastically different, many of their results are strikingly similar. Both studies found a more frequent occurrence of modals in speech than in writing. Although the rankings are different, both studies found that *will, can,* and *would* were the most frequently occurring modals in speech, and that *will* and *would* were the most frequent modals in writing. Certain modals tended to occur most frequently in one medium rather than the other: *may* in writing more often than speech; *shall* more often in speech than in writing. Even though both studies contain frequency information on the meanings of modals, it is difficult to make direct comparisons: the two studies used different categories to classify the meanings of modals, and Coates (1983) calculated frequencies based only on an analysis of one of her corpora (the London Corpus), biasing her results more towards speech and certain kinds of unprinted material. Nevertheless, the results that can be compared illustrate that frequently occurring grammatical constructions can be reliably studied in relatively small corpora.

1.3.2 Reference grammars

While it is quite common to use corpora to investigate a single grammatical construction in detail, it is also possible to use corpora to obtain information on the structure and usage of many different grammatical constructions and to use this information as the basis for writing a reference grammar of English.

As was noted in the Preface, there is a long tradition in English studies, dating back to the nineteenth and early twentieth centuries, to use some kind of corpus as the basis for writing a reference grammar of English, a tradition followed by grammarians such as Jespersen (1909–49) or Curme (1947), who based their grammars on written material taken from the works of eminent English writers. Many modern-day reference grammars follow this tradition, but instead of using the kinds of written texts that Jespersen and Curme used, have based their discussions of grammar on commonly available corpora of written and spoken English. One of the first major reference works to use corpora were the two grammars written by Quirk, Greenbaum, Leech, and Svartvik: *A Grammar of Contemporary English* (1972) and *A Comprehensive Grammar of the English Language* (1985). In many sections of these grammars, discussions of grammatical constructions were informed by analyses of the London Corpus. For instance, Quirk et al.'s (1985: 1351) description of the noun phrase concludes with a table presenting frequency information on the distribution of simple and complex noun phrases in various genres of the London Corpus. In this table, it is pointed out that in prose fiction and informal spoken English, a sentence with the structure of (9) would be the norm: the subject contains a simple noun phrase (the pronoun *he*) and the object a complex noun phrase consisting of a head noun (*guy*) followed by a relative clause (*who is supposed to have left*).

(9) He's the guy who is supposed to have left (ICE-GB S1A-008 266)

In scientific writing, in contrast, this distribution of simple and complex noun phrases was not as common. That is, in scientific writing, there was a greater tendency to find complex noun phrases in subject positions. Thus, in scientific writing, it was not uncommon to find sentences such as (10), a sentence in which a very complex noun phrase containing a head (*those*) followed by a relative clause (*who have . . .*) occurs in subject position of the sentence:

(10) Even those who have argued that established, traditional religions present a major hegemonic force can recognize their potential for developing an "internal pluralism." (ICE-GB:W2A-012 40)

Information of this nature is included in the Quirk et al. grammars because one of the principles underlying these grammars is that a complete description of English entails information not just on the form of grammatical constructions but on their use as well.

More recent reference grammars have relied even more heavily on corpora. Like the Quirk et al. grammars, these grammars use corpora to provide information on the form and use of grammatical constructions, but additionally contain extensive numbers of examples from corpora to illustrate the grammatical constructions under discussion. Greenbaum's *Oxford English Grammar* (1996) is based almost entirely on grammatical information extracted from the British Component of the International Corpus of English (ICE-GB). The Collins COBUILD Project has created a series of reference grammars for learners of English that contains examples drawn from Bank of English Corpus (Sinclair 1987). Biber et al.'s *Longman Grammar of Spoken and Written English* (1999) is based on the Longman Spoken and Written English Corpus, a corpus that is approximately 40 million words in length and contains samples of spoken and written British and American English. This grammar provides extensive information not just on the form of various English structures but on their frequency and usage in various genres of spoken and written English.

1.3.3 Lexicography

While studies of grammatical constructions can be reliably conducted on corpora of varying length, to obtain valid information on vocabulary items, 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 shorter corpora, such as the one-million-word LOB Corpus. In the LOB Corpus, the five most frequent lexical items are the function words *the, of, and, to,* and *a*. The five least frequent lexical items are not five single words but rather hundreds of different words that occur from ten to fifteen times each in the corpus. These words include numerous proper nouns as well as miscellaneous content words such as *alloy, beef,* and *bout*. These frequencies

illustrate a simple fact about English vocabulary (or, for that matter, vocabulary patterns in any language): a relatively small number of words (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 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 is the philosophy of the Collins COBUILD Project at Birmingham University in England, which has 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 Bank of English Corpus, which as of October 2000 totaled 415 million words.

The Bank of English Corpus has many potential uses, but it was designed primarily to help in the creation of dictionaries. Sections of the corpus were used as the basis of the *BBC English Dictionary*, a dictionary that was intended to reflect the type of vocabulary used in news broadcasts such as those on the BBC (Sinclair 1992). Consequently, the vocabulary included in the dictionary was based on sections of the Bank of English Corpus containing transcriptions of broadcasts on the BBC (70 million words) and on National Public Radio in Washington, DC (10 million words). The Bank of English Corpus was also used as the basis for a more general purpose dictionary, the *Collins COBUILD English Dictionary*, and a range of other dictionaries on such topics as idioms and phrasal verbs. Other projects have used similar corpora for other types of dictionaries. The Cambridge Language Survey has developed two corpora, the Cambridge International Corpus and the Cambridge Learners' Corpus, to assist in the writing of a number of dictionaries, including the *Cambridge International Dictionary of English*. Longman publishers assembled a large corpus of spoken and written American English to serve as the basis of the *Longman Dictionary of American English*, and used the British National Corpus as the basis of the *Longman Dictionary of Contemporary English*.

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 also 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 fifty years to complete, largely because of the many stages of production that the dictionary went through. Landau (1984: 69) notes that the 5 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 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 (cf. section 5.3.2), 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. In addition, some concordancing programs can detect prefixes and suffixes and irregular forms and sort words by "lemmas": words such as *runs, running,* and *ran* will not be counted as separate entries but rather as variable forms of the lemma *run*. To study the meanings of individual words, the lexicographer can have a word displayed in KWIC (key word in context) format, and easily view the varying contexts in which a word occurs and the meanings it has in these contexts. And 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; cf. section 4.3), the part of speech of each word can be automatically determined. In short, the computer corpus and associated software have completely revolutionized the creation of dictionaries.

In addition to making the process of creating a dictionary easier, corpora can improve the kinds of information about words contained in dictionaries, and address some of the deficiencies inherent in many dictionaries. One of the criticisms of the *OED*, Landau (1984: 71) notes, is that it contains relatively little information on scientific vocabulary. But as the *BBC English Dictionary* illustrates, if a truly "representative" corpus of a given kind of English is created

(in this case, broadcast English), it becomes quite possible to produce a dictionary of any type of English (cf. section 2.5 for a discussion of representativeness in corpus design). And with the vast amount of scientific English available in computerized form, it would now be relatively easy to create a dictionary of scientific English that is corpus-based.

Dictionaries have also been criticized for the unscientific manner in which they define words, a shortcoming that is obviously a consequence of the fact that many of the more traditional dictionaries were created during times when well-defined theories of lexical meaning did not exist. But this situation is changing as semanticists turn to corpora to develop theories of lexical meaning based on the use of words in real contexts. Working within the theory of "frame" semantics, Fillmore (1992: 39–45) analyzed the meaning of the word *risk* in a 25-million-word corpus of written English created by the American Publishing House for the Blind. Fillmore (1992: 40) began his analysis of *risk* in this corpus working from the assumption that all uses of *risk* fit into a general frame of meaning that "there is a probability, greater than zero and less than one, that something bad will happen to someone or something." Within this general frame were three "frame elements," i.e. differing variations on the main meaning of *risk,* depending upon whether the "risk" is not caused by "someone's action" (e.g. *if you stay here you risk getting shot*), whether the "risk" is due in some part to what is termed "the Protagonist's Deed" (e.g. *I had no idea when I stepped into that bar that I was risking my life*), or whether the "risk" results from "the Protagonist's decision to perform the Deed" (e.g. *I know I might lose everything, but what the hell, I'm going to risk this week's wages on my favorite horse*) (Fillmore 1992: 41–2).

In a survey of ten monolingual dictionaries, Fillmore (1992: 39–40) found great variation in the meanings of *risk* that were listed, with only two dictionaries distinguishing the three meanings of *risk*. In his examination of the 25-million-word corpus he was working with, Fillmore (1992) found that of 1,743 instances of *risk* he identified, most had one of the three meanings. However, there were some examples that did not fit into the *risk* frame, and it is these examples that Fillmore (1992: 43) finds significant, since without having examined a corpus, "we would not have thought of them on our own." Fillmore's (1992) analysis of the various meanings of the word *risk* in a corpus effectively illustrates the value of basing a dictionary on actual uses of a particular word. As Fillmore (1992: 39) correctly observes, "the citation slips the lexicographers observed were largely limited to examples that somebody happened to notice . . . " But by consulting a corpus, the lexicographer can be more confident that the results obtained more accurately reflect the actual meaning of a particular word.

1.3.4 Language variation

Much of the corpus-based research discussed so far in this section has described the use of either grammatical constructions or lexical items in some

kind of context: speech vs. writing, or scientific writing vs. broadcast journalism. The reasons these kinds of studies are so common is that modern-day corpora, from their inception, have been purposely designed to permit the study of what is termed "genre variation," i.e. how language usage varies according to the context in which it occurs. The first computer corpus, the Brown Corpus, contained various kinds of writing, and this corpus design has influenced the composition of most "balanced" corpora created since then.

Because corpus linguists have focused primarily on genre variation, they have a somewhat different conception of language variation than sociolinguists do. In sociolinguistics, the primary focus is 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 (a point that is discussed in greater detail in section 2.5). 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 subcorpus 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 complications that studying linguistic variables poses, designers of some recent corpora have made more concerted efforts to create corpora that are balanced for such variables as age and gender, and that are set up in a way that information on these variables can be extracted by various kinds of software programs. Prior to the collection of spontaneous dialogues in the British National Corpus, 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 subcorpus 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 different boroughs of London. To enable the study of sociolinguistic variables in the spoken part of the BNC, each conversation contains a file header (cf. section 4.1), 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, 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 (cf. section 5.3.2), can draw upon information in this database to restrict searches. Even though ICE-GB is not a balanced corpus – 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 use 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 it 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.

Although traditionally designed corpora such as Brown or LOB might seem deficient because they do not easily permit the study of sociolinguistic variables, this deficiency has been more than compensated for by the important information on genre variation that these corpora have yielded. Biber's (1988) study of the linguistic differences between speech and writing effectively illustrates the potential that corpora have for yielding significant insights into the structure 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 writing. To reach this conclusion, Biber (1988) first used a statistical test, factor analysis (explained in section 5.4.1), to determine which linguistic constructions tended to co-occur in particular texts. Biber (1988: 13) was interested in grammatical co-occurrences because he believes "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." High on this dimension are two types of written texts that are dense in linguistic constructions that are markers of abstract language: academic prose and official documents. Low on the dimension are two types of spoken texts that contain relatively few abstractions: face-to-face conversations and telephone conversations. However, Biber (1988) also found that certain kinds of written texts (e.g. romantic fiction) were low on the dimension, and certain kinds of spoken texts (e.g. prepared speeches) were higher on the dimension. Findings such as this led Biber (1988) to conclude that there is no absolute difference between speech and writing. More recently, Biber (1995) has extended this methodology to study genre analysis in corpora of languages other than English, and in corpora containing texts from earlier periods of English (Biber and Burges 2000).

1.3.5 Historical linguistics

The study of language variation is often viewed as an enterprise conducted only on corpora of Modern English. However, there exist a number of historical corpora – corpora containing samples of writing representing earlier dialects and periods of English – that can be used to study not only language variation in earlier periods of English but changes in the language from the past to the present.

Much of the interest in studying historical corpora stems from 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 subperiods (ranging from 70–100 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 and different genres (e.g. law, homilies and sermons, fiction, and letters; for full details, see Kytö 1996); and it 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 English Historical Registers) was created (cf. Biber et al. 1994; Biber and Burges 2000). ARCHER is 1.7 million words in length and contains American as well as British English. It covers the years 1650–1990, with texts divided into fifty-year subgroups. 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.

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 1961 and 1991. 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 thirty years. Although thirty 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, are what Rissanen (2000) terms "multi-purpose general corpora" because they 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 electronic versions of *Beowulf* ("The Electronic *Beowulf*"; cf. Prescott 1997 and http://www.uky.edu/~kiernan/eBeowulf/guide.htm); the works of Chaucer and other Middle English writers (the Corpus of Middle English Prose and Verse; cf. http://www.hti.umich.edu/english/mideng/); 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). As Rissanen's (2000) discussion of these and other historical corpora indicates, the creation of historical corpora has become as active an enterprise as the creation of corpora of Modern English.

Historical corpora have greatly enhanced our ability to study the linguistic development of English: such corpora allow corpus linguists not only to study systematically the development of particular grammatical categories in English but to gain insights into how genres in earlier periods differed linguistically and how sociolinguistic variables such as gender affected language usage.

Finegan and Biber's (1995) study of the alternation of *that* and zero in constructions such as *I know (that) this is true* provides a good example of the types of genre variation that can be studied in historical corpora. Finegan and Biber (1995) studied the alternation of *that* and zero in various genres of ARCHER and the Helsinki Corpus. In Modern English, this alternation is a marker of formal vs. informal style. As Greenbaum, Nelson and Weizman's (1996: 84) study of speech and writing demonstrated, *that* tended to be more common in formal styles, such as academic writing and monologues, and much less common in informal styles, such as conversations and personal letters. This trend, however, was only partially confirmed by Finegan and Biber (1995): although the most colloquial written genre that they investigated, personal letters, contained the most instances of zero (at least in some periods), this genre (and the two more written-based genres, sermons and medicine) showed that historically in written genres of English, *that* has been preferred over zero in complement clauses, and that the trend towards *that* has increased in all genres since 1750. Of course, these results are tentative, as Finegan and Biber (1995) correctly observe. Nevertheless, they demonstrate the potential that historical corpora have for documenting changes in the language in both written and speech-based genres of English.

The study of sociolinguistic variables in historical corpora has provided numerous insights into the use of language in earlier periods, particularly in the area of gender differences. Nevalainen (2000) used the Corpus of Early English Correspondence to explore a hypothesis concerning gender that is popular in sociolinguistics, namely that rather than simply "*favour*[ing] prestige forms," females "in fact *create* them" (emphases in original). Nevalainen (2000) chose personal letters as the basis of the study because they are the best kind of genre from earlier periods of English to study "the day-to-day interaction between people," the kind of interaction in which gender differences are likely to emerge. In her corpus, Nevalainen (2000) analyzed gender differences in the use of three different linguistic constructions: the replacement in subject positions of *ye* by *you*, the replacement of the third-person-singular verb suffix-*(e)th* by-*(e)s*, and the loss of multiple negation and subsequent increase in the use of non-assertive indefinites such as *any* or *ever*. While Nevalainen (2000) found that the first two changes were "led by female writers," the loss of double negation was "promoted by males," a change that Nevalainen (2000) attributes to "the two sexes' differential access to education in general and to professional specializations in particular." These mixed results lead Nevalainen (2000) to conclude that women's role in language change was different in the sixteenth century than it is in the twentieth century, and that further information is needed to determine precisely when women came to generally promote linguistic change.

1.3.6 Contrastive analysis and translation theory

Corpora, in particular "parallel" corpora, have been created to facilitate contrastive analyses of English and other languages, advance developments in translation theory, and enhance foreign language teaching.

One of the earlier parallel corpora, the English–Norwegian Parallel Corpus, provides a good illustration of how parallel corpora have been designed and of the particular kinds of linguistic analyses that can be conducted on them. The English–Norwegian Parallel Corpus contains samples of English and Norwegian fiction and non-fiction that are 10,000–15,000 words in length (Johansson and Ebeling 1996). This corpus was designed to permit three kinds of studies. First, because the corpus contains samples of similar types of fiction and non-fiction written originally in English or Norwegian, it can be used to study genre variation between English and Norwegian: how English fiction, for instance, is similar to or different than Norwegian fiction. Second, each sample has been translated from the original language into the other language. Consequently, the corpus allows for the study of how English is typically translated into Norwegian, or Norwegian into English. Finally, the corpus can be used to compare the structure of, for instance, fiction written originally in Norwegian with fiction translated into Norwegian, or non-fiction written originally in English with non-fiction translated into English.

To analyze the corpus, software has been developed to align sentences and to conduct various kinds of searches. For instance, one can search for an English expression such as *as it were* and receive a list of every sentence containing this expression as well as a sentence containing the Norwegian translation of the expression:

(11) <s id=FW1.4.s153 corresp=FW1T.4.s154>She took a swig of whiskey and tried to relocate herself, *as it were*, in her own skin.</s>
<s id=FW1T.4.s154 corresp=FW1.4.s153>Hun tok seg en god slurk med whisky, og prøvdeågjenfinne seg selv *liksom*, i sitt eget skinn.</s> (Johansson and Ebeling 1996: 6)

This kind of display permits a range of different contrastive studies to be conducted on the corpus. Hasselgård (1997) studied "sentence openings": the different grammatical constructions that English and Norwegian use to begin sentences. She found that in the original English and Norwegian samples there were distinct differences between the elements placed in the opening section of a sentence (e.g. Norwegian more frequently places direct objects at the beginning of a sentence than English does). However, in the translations, there were fewer differences, with "the source language" having a greater effect "on the syntax of the target language" (Hasselgård 1997: 18).

This influence that Hasselgård (1997) discovered in her study is commonly referred to as "translationese" (Johansson and Hofland 1993), i.e. features of the translated text more characteristic of the source language than the language the sentence is being translated into. Schmied and Schäffler (1996) describe a number of examples of translationese in a parallel corpus of English and German, the Chemnitz Corpus. They note that translationese can involve not just particular syntactic and morphological differences but pragmatic differences as well. For instance, in German, it is very common to express tentativeness with adverbials or modal particles. However, Schmied and Schäffler (1996: 50) found

a number of examples in their corpus tending more towards the English norm of using in the case of (12) a lexical modal structure in the German translation rather than the more natural modal particle (given in brackets at the end of the example).

(12) In the less developed areas it has *tended to* widen. (Commission of the European Union, 1993: 10)
[. . .] in den weniger entwickelten Teilen der Gemeinschaft bestand sogar die *Tendenz* zu einer Vergrösserung def Kluft. [natural language: hat sich die Kluft *eher* vergrössert] (Kommission der Europäischen Gemeinschaften, 1993: 10)

Schmied and Schäffler (1996: 52) observe that information on translation such as this can be used to train translators or help create bilingual dictionaries.

Because of the increased popularity of parallel corpora, there now exists general-purpose software, such as ParaConc (Barlow 1999), that can be used to align sentences in any two or more languages and to aid researchers in conducting the kinds of studies detailed in this section.

1.3.7 Natural language processing

The primary emphasis in this chapter has been on how corpora can be used to conduct linguistic analyses of English. However, there are many corpus linguists whose interests are more computational than linguistic. These linguists have created and used corpora to conduct research in an area of computational linguistics known as "natural language processing" (NLP). For instance, the North American Chapter of the Association for Computational Linguistics regularly has workshops and special sessions at which computational linguists in NLP discuss the use of corpora to advance research in such areas as tagging, parsing, information retrieval, and the development of speech recognition systems. One recent special session at the 2000 conference in Seattle, Washington, on "Large Corpus Annotation and Software Standards: Towards an American National Corpus" provided a forum for the discussion of the necessary software and annotation standards required for annotating very large corpora, such as the American National Corpus, which is intended to be the American equivalent of the British National Corpus (cf. chapter 4 for more details on corpus annotation).

Because researchers in NLP have their own distinct interests, the corpora they use are designed differently than corpora such as Brown or LOB. This point is illustrated by a survey of the corpora distributed by the Linguistic Data Consortium (LDC), an organization working out of the University of Pennsylvania (http://www.ldc.upenn.edu/). Many LDC corpora have been used to develop speech recognition systems. The TIMIT Acoustic–Phonetic Continuous Speech Corpus contains digitized recordings of 630 speakers, representing various dialect regions in the United States, who were recorded reading a series of sentences. Aligned with each speaker's recording are phonetic as well as standard

orthographic transcriptions of each word a speaker utters. The Switchboard Corpus contains digitized recordings of approximately 2,400 telephone conversations. The conversations are orthographically transcribed and the transcriptions are linked to the audio (cf. Wheatley et al. 1992). The TIMIT and Switchboard corpora are but two of the many LDC distributed corpora used to improve speech recognition systems.

Other LDC corpora have been used to develop taggers and parsers and information retrieval systems. The Penn Treebank contains a heterogeneous collection of spoken and written texts totaling 4.9 million words; sections of this corpus have been tagged and parsed (cf. Marcus, Santorini, and Marcinkiewicz 1993 and http://www.cis.upenn.edu/~treebank/cdrom2.html, and also section 4.5 for a more detailed discussion of treebanks). The TIPSTER Corpus contains a variety of written texts taken from such sources as the *Wall Street Journal* and Associated Press newswire stories. Because the corpus is used mainly to develop information retrieval systems, it is not tagged and parsed but annotated with SGML tags, which are designed to reveal the structure of documents rather than their linguistic structure (cf. section 4.1 for more information on SGML tags).

Corpora such as those described above have not been used that extensively by corpus linguists doing descriptive studies of English, largely because many of the corpora used by researchers in NLP are not suitable for many kinds of linguistic analyses. In the abstract, the kinds of texts used in the Switchboard Corpus – telephone conversations – would be very useful for linguistic analysis. However, conversants whose speech is included in this corpus were instructed beforehand to discuss a specific topic. Therefore, the speech in this corpus is not "natural" (cf. section 3.2 for a discussion of natural speech). The Penn Treebank is not a balanced corpus: a large percentage of it is press reportage. Thus, the results of any linguistic analysis of this corpus will be limited to making generalizations about a very small number of genres.

But despite the limitations of the corpora created by researchers in NLP, the tools they have developed – taggers and parsers in particular – have been instrumental in creating grammatically analyzed corpora that are of great value to descriptive linguists. For instance, work done by computational linguists in developing the Nijmegen tagger and parser for the International Corpus of English greatly facilitated the grammatical annotation of the British component of ICE, and led to a fully tagged and parsed one-million-word corpus of spoken and written English that, when used with a text retrieval program (ICECUP), can extract from ICE-GB a wealth of grammatical information that in the past could be obtained only by manual analysis (cf. section 5.3.2 for more information on ICE-GB and ICECUP). The UCREL team at Lancaster University consists of a group of descriptive and computational linguists who worked together not only to create the British National Corpus but to develop the tagger (CLAWS) that was used to tag the corpus (Garside and Smith 1997). Speech recognition programs have progressed to the point where they can recognize with varying

degrees of accuracy not just carefully scripted monologues but broadcast speech (Nguyen et al. 1999). As these programs improve, they may in the future be used to at least partially automate the process of transcribing speech for inclusion in a corpus (cf. sections 3.5 and 3.6 for more on the transcription of speech). As these research initiatives illustrate, while their goals may be different, corpus linguists of all persuasions have much to contribute to each other's work.

1.3.8 Language acquisition

Although studies of language acquisition have always had an empirical basis, researchers in the areas of first- and second-language acquisition have tended not to make publicly available the data upon which their studies were based. However, this situation is changing and there now exist corpora suitable for studying both first- and second-language acquisition.

To facilitate the study of both first- and second-language acquisition, the CHILDES (Child Language Data Exchange) System was developed. This system contains a corpus of transcriptions of children and adults learning first and second languages that are annotated in a specific format called "CHAT" and that can be analyzed with a series of software programs called "CLAN" (MacWhinney 1996: 2). Although much of the corpus is focused on English, a considerable part of it represents nineteen additional languages as well, and the speech that is included is taken from normally developing individuals as well as those with language disorders such as aphasia or autism (MacWhinney 1996: 3). The CHILDES corpus has generated an impressive amount of research on language acquisition, ranging from studies of children learning Germanic and Romance languages to studies of children who have language disabilities (MacWhinney 2000: 421f.).

To study second-language acquisition, a number of researchers have begun developing what are called "learner corpora": corpora containing the speech or writing of individuals learning English as a second or foreign language. One of the larger corpora in this area is called the International Corpus of Learner English (ICLE). ICLE is currently more than two million words in length (http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/introduction.html); it is divided into 500-word essays written by students from fourteen different linguistic backgrounds learning English as a foreign language (Granger 1998: 10). Other learner corpora include the Longman's Learner Corpus (http://www.longman-elt.com/dictionaries/corpus/lclearn.html) and the Hong Kong University of Science and Technology (HKUST) Learner Corpus (Milton and Freeman 1996).

The creation of learner corpora has led to a number of studies in the area of what Granger (1998: 12–14) terms "contrastive interlanguage analysis (CIA)": the use of corpora to compare the interlanguage a native speaker of, for instance, French develops when learning English, and to compare across language groups the structure of the various interlanguages that individuals from different first-language backgrounds develop. Altenberg and Tapper (1998) employed this

type of contrastive interlanguage analysis in their study of the use of adverbial connectors in the section of ICLE containing the writing of Swedish students learning English (a subcomponent called "SWICLE"). Additionally, they compared the use of adverbial connectors by Swedish students with native-speaker usage. Although their conclusions are tentative, they found that overall the Swedish learners used connectives similarly to the way that native speakers do. Compared with French learners of English, however, the Swedish learners underused connectives, setting them apart from other language groups, which tend to overuse connectives. What led to this underuse, Altenberg and Tapper (1998: 92) note, was the tendency of Swedish learners to substitute less formal connectives for more formal connectives.

1.3.9 Language pedagogy

One consequence of the development of learner corpora is that researchers are taking information from them to develop teaching strategies for individuals learning English as a second or foreign language. After discovering that Swedish learners of English overused informal connectives in their writing, Altenberg and Tapper (1998: 92) concluded that the best strategy for teaching them the use of connectives would be to expose them "to a greater range of registers and to a more extensive training in expository writing." Gillard and Gadsby (1998) describe how they used the Longman's Learner Corpus (LLC) to help create a dictionary, the *Longman Essential Activator* (1997), designed specifically for students learning English. One way they used the corpus was to discover which errors students tended to make and then to use this information to design special "help boxes" included in the dictionary. For instance, Gillard and Gadsby (1998: 167) discovered many errors in the usage of the word *peoples* in the LLC, and they used this information to provide special information on the use of this word in its entry in the dictionary. The idea behind using learner corpora to develop teaching strategies is that they give teachers an accurate depiction of how their students are actually using the language, information that can then be incorporated into textbooks and lesson plans.

In addition to using information from learner corpora to develop teaching strategies for learners of English, many have advocated that students themselves study corpora to help them learn about English, a methodology known as "data-driven learning" (Johns 1994 and Hadley 1997). This method of teaching has students investigate a corpus of native-speaker speech or writing with a concordancing program to give them real examples of language usage rather than the contrived examples often found in grammar books. Kettemann (1995), for instance, provides examples of how students can use a concordancing program to investigate various kinds of grammatical distinctions in English, such as the difference between the past and present perfect. Kettemann (1995: 4) argues that a concordancing program is "an extremely powerful hypothesis testing device" that allows the student to conduct inductive explorations of grammatical

constructions "on vast amounts of data." Because concordancing programs can generate so much data, Gavioli (1997: 84) argues that students need to be taught how "to interpret the data" that they discover, a process that she claims can lead to an effective "language-learning activity."

The notion of data-driven learning has been extended to involve the ultimate corpus: the World Wide Web. The Lingua Center within the Division of English as an International Language at the University of Illinois has developed a methodology called the "Grammar Safari" (http://deil.lang.uiuc.edu/web.pages/grammarsafari.html). This methodology has students use search engines to find and then study grammatical constructions in various online resources, such as novels, newspapers, and magazines. Meyer (1998) describes the actual process that students can follow to find grammatical constructions on the WWW. He shows how a search engine, such as AltaVista, and a web browser, such as Netscape, can be used to study the disputed usage *between you and I* and to find not only examples of the usage but discussions of the usage in the archives of various e-mail discussion lists, such as the Linguist List and Ask a Linguist.

1.4 Conclusions

This chapter has demonstrated that corpora have numerous uses, ranging from the theoretical to the practical, making them valuable resources for descriptive, theoretical, and applied discussions of language. Because corpus linguistics is a methodology, all linguists – even generativists – could in principle use corpora in their studies of language. However, most generativists feel that a corpus enables one to study performance, not competence; as a result, they continue to use introspection as the primary source of their data. In many other disciplines of linguistics, corpora have proven to be valuable resources: they are used for creating dictionaries, studying language change and variation, understanding the process of language acquisition, and improving foreign- and second-language instruction. However, for corpora to be most effective, they must be compiled in a manner that ensures that they will be most useful to those who are potential users of them. This topic – how to plan, create, and annotate a corpus – is the subject of the next three chapters of the book.

Study questions

1. Corpus linguistics is based on the principle that linguistic theory and description is best based on real rather than contrived data. What are the advantages and disadvantages of obtaining data from a corpus rather than from introspection?
2. Why have lexicographers in recent years been using corpora to create dictionaries? What kinds of information can lexicographers get from corpora that they cannot get using more traditional approaches to dictionary creation?

3. Why is it easier to study genre variation in corpora than it is to study sociolinguistic variables, such as age and gender?
4. Why have language teachers started to use corpora to teach English as a second or foreign language? What can a student get from a corpus that he or she cannot get from a traditional book on grammar?

2 Planning the construction of a corpus

Before the texts to be included in a corpus are collected, annotated, and analyzed, it is important to plan the construction of the corpus carefully: what size it will be, what types of texts will be included in it, and what population will be sampled to supply the texts that will comprise the corpus. Ultimately, decisions concerning the composition of a corpus will be determined by the planned uses of the corpus. If, for instance, the corpus is be used primarily for grammatical analysis (e.g. the analysis of relative clauses or the structure of noun phrases), it can consist simply of text excerpts rather than complete texts. On the other hand, if the corpus is intended to permit the study of discourse features, then it will have to contain complete texts.

Deciding how lengthy text samples within a corpus should be is but one of the many methodological considerations that must be addressed before one begins collecting data for inclusion in a corpus. To explore the process of planning a corpus, this chapter will consider the methodological assumptions that guided the compilation of the British National Corpus. Examining the British National Corpus reveals how current corpus planners have overcome the methodological deficiencies of earlier corpora, and raises more general methodological considerations that anyone planning to create a corpus must 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 largest corpora ever created. Most of the corpus (about 90 percent) consists of various types of written British English, with the remainder (about 10 percent) comprised of different types of spoken British English. Even though writing dominates in the BNC, the amount of speech in the corpus is the most ever made available in a corpus.[1]

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

1. Even though the corpus would contain both speech and writing, more writing would be collected than speech.

[1] Prior to the creation of the British National Corpus, the London–Lund Corpus contained the most speech available in a corpus: approximately 500,000 words (see Svartvik 1990).

Table 2.1 *The composition of the British National Corpus (adapted from Aston and Burnard 1998: 29–33 and http://info.ox.ac.uk/bnc/what/balance.html)*

Speech			
Type	Number of texts	Number of words	% of spoken corpus
Demographically sampled	153	4,211,216	41%
Educational	144	1,265,318	12%
Business	136	1,321,844	13%
Institutional	241	1,345,694	13%
Leisure	187	1,459,419	14%
Unclassified	54	761,973	7%
Total	915	10,365,464	100%
Writing			
Type	Number of texts	Number of words	% of written corpus
Imaginative	625	19,664,309	22%
Natural science	144	3,752,659	4%
Applied science	364	7,369,290	8%
Social science	510	13,290,441	15%
World affairs	453	16,507,399	18%
Commerce	284	7,118,321	8%
Arts	259	7,253,846	8%
Belief & thought	146	3,053,672	3%
Leisure	374	9,990,080	11%
Unclassified	50	1,740,527	2%
Total	3,209	89,740,544	99%[2]

2. A variety of different genres of speech and writing would be gathered for inclusion in the corpus.
3. Each genre would be divided into text samples, and each sample would not exceed 40,000 words in length.
4. A number of variables would be controlled for in the entire corpus, such as the age and gender of speakers and writers.
5. 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; and who their target audience was.
6. 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.

[2] Because fractions were rounded up or down to the nearest whole number, cumulative frequencies in this and subsequent tables do not always add up to exactly 100 percent.

Even though the creators of the BNC had a very definite plan for the composition of the corpus, it is important to realize, as Biber (1993: 256) observes, that the creation of a corpus is a "cyclical" process, requiring constant re-evaluation as the corpus is being compiled. Consequently, the compiler of a corpus should be willing to change his or her initial corpus design if circumstances arise requiring such changes to be made. For instance, in planning the part of the BNC containing demographically sampled speech, it was initially thought that obtaining samples of speech from 100 individuals would provide a sufficiently representative sample of this type of speech (Crowdy 1993: 259). Ultimately, however, the speech of 124 individuals was required to create a corpus balanced by age, gender, social class, and region of origin (Aston and Burnard 1998: 32).

While changes in corpus design are inevitable, it is crucial that a number of decisions concerning the composition of a corpus are made prior to the collection of texts for the corpus, and the remainder of this chapter considers the factors that influenced the decisions detailed above that the creators of the BNC made, and the general methodological issues that these decisions raise.

2.2 The overall length of a corpus

Earlier corpora, such as Brown and LOB, were relatively short, primarily 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. This process has been greatly eased with the invention of optical scanners (cf. section 3.7), which can automatically computerize printed texts with fairly high rates of accuracy. However, technology has not progressed to the point where it can greatly expedite the collection and transcription of speech: there is much work involved in going out and recording speech (cf. section 3.2), and the transcription of spoken texts still has to be done manually with either a cassette transcription machine or special software that can replay segments of speech until the segments are accurately transcribed (cf. section 3.6). These logistical realities explain why 90 percent of the BNC is writing and only 10 percent speech.

To determine how long a corpus should be, it is first of all important to compare the resources that will be available to create it (e.g. funding, research assistants, computing facilities) with the amount of time it will take to collect texts for inclusion, computerize them, annotate them, and tag and parse them. Chafe, Du Bois, and Thompson (1991: 70–1) calculate that it will require "six person-hours" to transcribe one minute of speech for inclusion in the Santa Barbara Corpus of Spoken American English, a corpus containing not just orthographic transcriptions of speech but prosodic transcriptions as well (cf. section 4.1). For the American component of ICE, it was calculated that it would take eight hours of work to process a typical 2,000-word sample of

writing, and ten to twenty hours to process a 2,000-word spoken text (with spontaneous dialogues containing numerous speech overlaps requiring more time than carefully articulated monologues). Since the American component of ICE contains 300 samples of speech and 200 samples of writing, completing the entire corpus would take a full-time research assistant working forty hours a week more than three years to complete. If one applies these calculations to the BNC, which is one hundred times the length of the American component of ICE, it becomes obvious that creating a corpus requires a large commitment of resources. In fact, the creation of the BNC required the collaboration of a number of universities and publishers in Great Britain, as well as considerable financial resources from funding agencies.

After determining how large a corpus one's resources will permit, the next step is to consider how long the corpus needs to be to permit the kinds of studies one envisions for it. One reason that the BNC was so long was that one planned use of the corpus was to create dictionaries, an enterprise requiring a much larger database than is available in shorter corpora, such as Brown or LOB. On the other hand, all of the regional components of the ICE Corpus are only one million words in length, since the goal of the ICE project is not to permit in-depth lexical studies but grammatical analyses of the different regional varieties of English – a task that can be accomplished in a shorter corpus.

In general, the lengthier the corpus, the better. However, it is possible to estimate the required size of a corpus more precisely by using statistical formulas that take the frequency with which linguistic constructions are likely to occur in text samples and calculate how large the corpus will have to be to study the distribution of the constructions validly. Biber (1993: 253–4) carried out such a study (based on information in Biber 1988) on 481 text samples that occurred in twenty-three 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": 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 corpora).[3]

Unfortunately, such calculations presuppose that one knows precisely what linguistic constructions will be studied in a corpus. The ultimate length of a

[3] Sánchez and Cantos (1997) describe a formula they developed, based on a corpus of Spanish they created, to predict the length of corpus necessary to provide a reliable and accurate representation of Spanish vocabulary. However, because their calculations are based on the Spanish language, they caution that "future research" (p. 265) needs to be conducted to determine whether the formula can be applied to other languages, such as English, as well.

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.

2.3 The types of genres to include in 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) were "demographically sampled": they consisted of a variety of spontaneous dialogues and monologues recorded from individuals living in various parts of Great Britain.[4] 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) were "imaginative": 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 a number of 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), it turns out that while the two corpora contain the same range of genres, the genres are much more specifically delineated in the ICE Corpus (see table 2.2) than they 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 Corpus, this division is clearly reflected in the types of genres making up the spoken part of the corpus. 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 Corpus (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 Corpus, 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

[4] The types of individuals whose speech is represented in the demographically sampled part of the BNC are discussed in greater detail in section 2.5.

Table 2.2 *Composition of the ICE (adapted from Greenbaum 1996a: 29–30)*

Speech			
Type	Number of texts	Length	% of spoken corpus
Dialogues	180	360,000	59%
Private	100	200,000	33%
direct conversations	90	180,000	30%
distanced conversations	10	20,000	3%
Public	80	160,000	26%
class lessons	20	40,000	7%
broadcast discussions	20	40,000	7%
broadcast interviews	10	20,000	3%
parliamentary debates	10	20,000	3%
legal cross-examinations	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	60,000	10%
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
Non-printed	50	100,000	26%
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): humanities, social sciences, natural sciences, technology	40	80,000	20%
informational (popular): humanities, social sciences, natural sciences, technology	40	80,000	20%
informational (reportage)	20	40,000	10%
instructional: administrative, regulatory, skills, hobbies	20	40,000	10%
persuasive (press editorials)	10	20,000	5%
creative (novels, stories)	20	40,000	10%
Total	200	400,000	101%

type of writing in the BNC, in the ICE Corpus 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 Corpus, does not include equal numbers of texts in each of these categories. The ICE Corpus also contains a fairly significant number (25 percent) of non-printed 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 Corpus 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?

This question can be answered by considering the two main types of corpora that have been created, and the particular types of studies that can be carried out on them. The ICE Corpus 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 Corpus 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 Corpus 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 Corpus were collected). The ICE Corpus 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 Corpus.

Because general-purpose corpora such as the ICE Corpus and BNC do not always contain a full representation of a genre, it is now becoming quite common to see *special-purpose corpora* being developed. 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). By restricting the scope of the corpus, energy can be directed towards assembling a detailed collection of texts that fully represent the kind of academic language one is likely to encounter at an American university. As of June 2001, seventy-one texts (813,684 words) could be searched on the MICASE website: http://www.hti.umich.edu/micase/.

At the opposite extreme of special-purpose corpora like MICASE are those which have specialized uses but not for genre studies. The Penn Treebank consists of a heterogeneous collection of texts totalling 100 million words, ranging from Dow Jones newswire articles to selections from the King James version of the Bible (Marcus, Santorini, and Marcinkiewicz 1993). 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 "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 (cf. section 4.5 for more on parsers). And to accomplish this goal, all that is important is that a considerable number of texts are 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 (cf. section 3.2), for instance, or obtaining copyright clearance for written material (cf. section 3.3). 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 Helsinki, which contains a range of different genres (sermons, travelogues, fiction, drama, etc.; cf. section 1.3.5 for more details), as well as specialized corpora, such as the Corpus of Early English Correspondence, a corpus of letters written during the Middle English period. In gathering material for corpora such as these, the corpus compiler must deal with 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. Correspondence in the form of personal letters is not available until the fifteenth century. 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.

2.4 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. The Lampeter Corpus of Early Modern English Tracts consists of complete rather than composite texts. In the Helsinki Corpus, text samples range from 2,000–10,000 words in length. Samples within the BNC are equally varied but are as long as 40,000 words. This is considerably lengthier than earlier corpora, such as Brown or LOB, which contained 2,000-word samples, and the London–Lund Corpus, which contained 5,000-word samples. The ICE Corpus follows in the tradition of Brown and LOB and contains 2,000-word samples.[5]

Because most corpora contain relatively short stretches of text, they tend to contain text fragments rather than complete texts. 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. Experience with the creation of the ICE Corpus has shown that, in general, it is quite difficult to obtain permission to use copyrighted material, even if only part of a text is requested, and this difficulty will only be compounded if one seeks permission to use an entire text (cf. section 3.3).

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. In the American component of ICE (and many earlier corpora as well), sections of spoken texts are included that form coherent conversations themselves: many conversations consist of subsections that form coherent units, and that have their own beginnings, middles, and ends. Likewise, for written texts, one can include

[5] All of these lengths are approximate. For instance, some texts in the ICE Corpus exceed 2,000 words to avoid truncating a text sample in mid-word or mid-paragraph.

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 Corpus 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 totalling 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. He 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. He 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 thousand words of a newspaper article totalling 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 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 why 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.[6]

2.5 Determining the number of texts and range of 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 a 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 considerations that need to be taken into account 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 Corpus 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 Corpus felt that spontaneous dialogues are a very important type of spoken English and should therefore be

[6] It is also possible to use elicitation tests to study infrequently occurring grammatical items – tests that ask native speakers to comment directly on particular linguistic items. See Greenbaum (1973) for a discussion of how elicitation tests can be used to supplement corpus studies.

amply represented. The reason for this sentiment is obvious: while only a small segment of the speakers of English create 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% conversation and 3% letters and notes, with the remaining 7% 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 subgenres, the subgenres in official documents (e.g. government reports, business reports, and university documents) are much more linguistically similar than the subgenres 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 subgenres 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 subgenres 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 very lengthy corpus, it provides a sufficient number of samples of genres to enable generalizations to be made about the genres. However, with the much shorter ICE Corpus (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 Corpus, 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 (1992) used 120,000-word sections of the Brown, the London–Lund, and the Survey of English Usage (SEU) Corpora to study the usage of appositions in four genres: spontaneous conversation, press reportage, academic prose, and fiction. In analyzing only twenty 2,000-word samples from the press genre of the Brown Corpus, he found ninety-seven examples of appositions of the type *political neophyte Steve Forbes* in which the first unit lacks a determiner and occurs before a proper noun (p. 12); in eight 5,000-word samples from the press genre of the SEU Corpus, he found only thirty-one examples of such appositions. These findings allowed Meyer (1992) to claim quite confidently that not only were such appositions confined to the press genre (the other genres contained virtually no examples of this type of apposition) but they were more common in American English than in British English.

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. Although sampling methodology can be used to determine the number of samples needed to represent a genre, this methodology, as was demonstrated earlier in this section, is not all that useful for this purpose. It is best used to select the individuals whose speech and writing will be included in a corpus.

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 creating the Tampere Corpus, Norri and Kytö (1996) decided that their sampling frame would not be just scientific writing but different types of scientific writing written at varying levels of technicality. After deciding what their sampling frame was, Norri and Kytö (1996) then had to determine which elements from this frame they needed to include in their corpus to produce a representative sampling of scientific writing, that is, enough writing from the social sciences, for instance, so that someone using the corpus could analyze the writing and be able to claim validly that their results were true not just about the specific examples of social science writing included in the Tampere Corpus but all social science writing. Obviously, Norri and Kytö (1996) could not include all examples of social science writing published in, say, a given year; they had to therefore narrow the range of texts that they selected for inclusion.

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 2,168–2,289 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 and 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 twelve 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 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 non-probability 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 before-hand 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, one 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. 10 percent African Americans, 15 percent Hispanic Americans, etc.).

Although probability sampling is the most reliable type of sampling, leading to the least amount of bias, for the corpus linguist this kind of sampling presents considerable logistical challenges. The mathematical formulas used

in probability sampling often produce very large sample sizes, as the example with books cited above illustrated. Moreover, to utilize the sampling techniques employed by creators of the BNC in the United States would require considerable resources and funding, given the size of the United States and the number of speakers of American English. Consequently, it is quite common for corpus linguists to use non-probability sampling techniques in compiling their corpora.

In creating the Brown Corpus, for instance, a type of "judgement" sampling was used. That is, prior to the creation to 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:

(a) for newspapers, the microfilm files of the New York Public Library;
(b) for detective and romantic fiction, the holdings of the Providence Athenaeum, a private library;
(c) for various ephemeral and popular periodical material, the stock of a large secondhand magazine store in New York City;
(d) 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 the International Corpus of English used a combination of "judgement" 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 Corpus of Spoken American English has 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.

The discussion thus far in this chapter has focused on the composition of the BNC to identify the most important factors that need to be considered prior to collecting texts for inclusion in a corpus. This discussion has demonstrated that:

1. Lengthier corpora are better than shorter corpora. However, even more important than the sheer length of a corpus is the range of genres included within it.
2. The range of genres to be included in a corpus is determined by whether it will be a multi-purpose corpus (a corpus intended to have wide uses) or a special-purpose corpus (a corpus intended for more specific uses, such as the analysis of a particular genre like scientific writing).

3. It is more practical to include text fragments in a corpus rather than complete texts. These fragments can be as short as 2,000 words, especially if the focus of study is frequently occurring grammatical constructions.
4. The number of samples of a genre needed to ensure valid studies of the genre is best determined by how much internal variation exists in the genre: the more variation, the more samples needed.
5. Probability sampling techniques can also be used to determine the number of samples of a genre necessary for inclusion in a corpus. However, such techniques are better used for selecting the individuals from whom text samples will be taken.
6. If it is not possible to select individuals using probability sampling techniques, then non-probability sampling techniques can be used, provided that a number of variables are considered and controlled for as well as possible.

These variables will be the focus of the remainder of this chapter.

2.6 The time-frame for selecting speakers and 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. Although most of the written texts used in the BNC were written or published between 1975 and 1993, a few texts date back as far as 1960 (Aston and Burnard 1998: 30). The Brown and LOB corpora contain texts published in 1961. The ICE Corpus contains texts spoken or written (or published) in the years 1990-present.

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. For instance, Mair (1995) analyzed the types of verb complements that followed the verb *help* in parts of the press sections of the LOB Corpus (comprised of texts published in 1961) and the FLOB (Freiburg–Lancaster–Oslo–Bergen) Corpus, a corpus comparable to LOB but containing texts published thirty years later in 1991. Mair (1995) found some differences in the complementation patterns, even over a period as short as thirty years. While *help* with the infinitive *to* was the norm in the 1961 LOB Corpus (e.g. *I will help you to lift the books*), this construction has been replaced in the 1991 FLOB Corpus by the bare infinitive (e.g. *I will help you lift the books*).

Mair's (1995) study indicates that language change can occur over a relatively short period of time. Therefore, if one wishes to create a truly synchronic corpus,

then the texts included in the corpus should reflect as narrow a time-frame as possible. Although this time-frame does not need to be as narrow as the one-year frame in Brown and LOB, a time-frame of five to ten years seems reasonable.

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 subperiods, the Late Middle and Early Modern English periods into seventy- to eighty-year periods. The lengths of subperiods were shorter in the later periods of the corpus "to include the crucial decades of the gradual formation of the fifteenth-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 predetermined 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.

2.7 Sampling native vs. non-native speakers of English

If one is setting out to create a corpus of American English, it is best to include only native speakers of American English in the corpus, since the United States is a country in which English is predominantly a language spoken by native speakers of English. On the other hand, if the goal is to create a corpus of Nigerian English, it makes little sense to include in the corpus native speakers of English residing in Nigeria, largely because Nigerian English is a variety of English spoken primarily as a second (or additional) language. As these two types of English illustrate, the point is not simply whether one obtains texts from native or non-native speakers but rather that the texts selected for inclusion are obtained from individuals who accurately reflect actual users of the particular language variety that will make up the corpus.

Prior to selecting texts for inclusion in a corpus, it is crucial to establish criteria to be used for selecting the individuals whose speech or writing will be included in the corpus. Since the American component of ICE was to include native speakers of American English, criteria were established precisely defining a native speaker of American English. Although language theorists will differ in their definitions of a native speaker of a language, for the American component of ICE, a native speaker of American English was defined as someone who

had lived in the United States and spoken American English since adolescence. Adolescence (10–12 years of age) was chosen as the cut-off point, because when acquisition of a language occurs after this age, foreign accents tend to appear (one marker of non-native speaker status). In non-native varieties of English, the level of fluency among speakers will vary considerably, and 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 define 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 keep track of the sociolinguistic variables described in section 2.8), 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 the American component of ICE, individuals are asked to supply their residence history: the places they have lived over their lives and the dates they have lived there. If, for instance, an individual has spent the first four years of life living outside the United States, it can probably be safely assumed that the individual is a native speaker of English and came to the United States at an early enough age to be considered a native speaker of English. On the other hand, if an individual spent the first fifteen years of life outside the United States, he or she is probably a non-native speaker of English and has learned English as a second language. It is generally best not to ask individuals directly on the biographical form whether they are a native speaker or not, since they may not understand (in the theoretical sense) precisely what a native speaker of a language is. 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. When dealing with written texts from earlier periods, additional problems can be encountered: English texts were often translations from Latin or French, and a text published in the United States, for instance, might have been written by a speaker of British English. Ultimately, however, when dealing with written texts, one simply has to 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 native speakers of the variety being collected. However, this does not necessarily mean that the text has to 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 (cf. section 4.1), that is, material not considered part of the corpus for purposes of word counts, generating KWIC (key word in context) concordances, and so forth.

2.8 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, a number of 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 on a corpus that these variables will have are summarized below.

2.8.1 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 can be requested on a biographical form, and in written texts, one can usually 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 of the variables affecting the gender balance of a corpus. The best the corpus compiler can do is be aware of the variables, confront them head on, and deal with them as much as is possible during the construction of a corpus.

2.8.2 Age

Although there are special-purpose corpora such as CHILDES (cf. section 1.3.8) that contain the language of children, adolescents, and adults, most corpora contain the speech and writing of "adults." The notable exception to this trend is the British National Corpus. In the spoken section of the corpus containing demographically sampled speech, there is a balanced grouping of texts representing various age groups, ranging from the youngest grouping (0–14 years) to the oldest grouping (60+) ("Composition of the BNC": http://info.ox.ac.uk/bnc/what/balance.html). In addition, there were texts included in the BNC taken from the Bergen Corpus of London Teenager English (COLT), which contains the speech of adolescents up to age 16 (Aston and Burnard 1998: 32; also cf. section 1.3.4). However, the written part of the BNC contains a sparser selection of texts from younger age groups, largely because younger individuals simply do not produce the kinds of written texts (e.g. press reportage or technical reports) represented in the genres typically included in corpora. This is one reason why corpora have tended to contain

mainly adult language. Another reason is that to collect the speech of children and adolescents, one often has to 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.

If it is decided that a corpus is to contain only adult language, it then becomes necessary to determine some kind of age cut-off for adult speech. The ICE project has set the age of 18 as a cut-off point between adolescent and adult, and thus has included in the main body of the corpus only the speech and writing of individuals over the age of 18. However, even though one may decide to include only adult speech in a corpus, to get a full range of adult speaking styles, it is desirable to include adults conversing with adolescents or children as well. In the transcribed corpus itself, markup can be included to set off the speech of adolescents or children as "extra corpus" material.

In most corpora attempting to collect texts from individuals of varying ages, the ages of those in the corpus are collapsed into various age groups. For instance, the British component of ICE has five groups: 18–25, 26–45, 46–65, 66+, and a category of "co-authored" for written texts with multiple authors.

2.8.3 Level of education

It is also important to consider the level of education of those whose speech or writing is to be included in the corpus. One of the earlier corpora, the London Corpus, set as its goal the collection of speech and writing from *educated* speakers of British English. This goal presupposes not only that one can define an educated speaker but that it is desirable to include only the language of this group in a corpus. The ICE project defined an educated speaker very precisely, restricting texts included in all of the components of ICE to those with at least a high school education. Arguably, such a restriction is arbitrary and excludes from the ICE Corpus a significant range of speakers whose language is a part of what we consider Modern English. Moreover, restricting a corpus to educated speech and writing is elitist and seems to imply that only educated speakers are capable of producing legitimate instances of Modern English.

However, even though these are all valid criticisms, it is methodologically valid to restrict a corpus to the speech and writing of whatever level of education one wishes, provided that no claims are made that such a corpus is truly representative of all speakers of a language. Consequently, the more representative one wishes a corpus to be, the more diverse the educational levels one will want to include in the corpus. In fact, a corpus compiler may decide to place no restrictions on the education level of those to be included in a corpus. However, if this approach is taken, it is important to keep careful records of the educational levels of those included in the corpus, since research has shown that language usage varies by educational level, and future users of the corpus may wish to use it to investigate language usage by level of education.

2.8.4 Dialect variation

The issue of the validity of including only educated speech in a corpus raises a more general consideration, namely the extent to which a corpus should contain the 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." In addition, more detailed information on writers is really not available until the fifteenth century. 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 an enormous 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 subdialects (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 British National Corpus, twelve dialect regions were identified in Great Britain, and from these dialect regions, 124 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 124 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, it is more profitable for corpus linguists interested in studying social and regional variation to devote 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 containing transcriptions of interviews of speakers of Hiberno English. Once a number of more focused corpora like these are created, they can be compared with one another and valid comparisons of larger dialect areas can then be conducted.

2.8.5 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 that both the social context in which speech occurs and the social relationships between speakers have a tremendous influence on the structure of speech. 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 American component of ICE contains spontaneous conversations taking place in many different social contexts: there are recordings of family members conversing over dinner, friends engaging in informal conversation as they drive in a car, co-workers speaking at work about work-related matters, teachers and their students discussing class work at a university, individuals talking over the phone, and so forth. The Michigan Corpus of Academic Spoken English (MICASE) collected academic speech in just about every conceivable academic context, from lectures given by professors to students conversing in study groups, to ensure that the ultimate corpus created represented 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.

2.9 Conclusions

To create a valid and representative corpus, it is important, as this chapter has shown, to plan the construction carefully 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 vs. 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 valid for the particular analysis being conducted. This shared responsibility will ensure that corpora become the most effective tools possible for linguistic research.

Study questions

1. Is a lengthier corpus necessarily better than a shorter corpus? What kinds of linguistic features can be studied in a lengthier corpus (such as the Bank

of English Corpus) that cannot be studied in a shorter corpus (such as the Brown Corpus)?
2. Would a corpus containing ten 2,000-word samples of writing from the humanities be lengthy enough to provide a representative sampling of humanistic writing?
3. Why is achieving gender balance in a corpus more than a matter of simply selecting texts from an equal number of males and females?
4. Why can FLOB (the Freiburg London–Oslo–Bergen Corpus) and FROWN (the Freiburg–Brown Corpus) be considered synchronic as well as diachronic corpora?
5. If the goal is to create a corpus to study regional or social variation, why does it make sense to target a very specialized regional or social dialect rather than a larger one, such as British English or American English?

3 Collecting and computerizing data

Once the basic outlines of a corpus are determined, it is time to begin the actual creation of the corpus. This is a three-part process, involving the collection, computerization, and annotation of data. This chapter will focus on the first two parts of this process – how to collect and computerize data. The next chapter will focus in detail on the last part of the process: the annotation of a corpus once it has been encoded into computer-readable form.

Collecting data involves recording speech, gathering written texts, obtaining permission from speakers and writers to use their texts, and keeping careful records about the texts collected and the individuals from whom they were obtained. How these collected data are computerized depends upon whether the data are spoken or written. Recordings of speech need to be manually transcribed using either a special cassette tape recorder that can automatically replay segments of a recording, or software that can do the equivalent with a sample of speech that has been converted into digital form. Written texts that are not available in electronic form can be computerized with an optical scanner and accompanying OCR (optical character recognition) software, or (less desirably) they can be retyped manually.

Even though the process of collecting, computerizing, and annotating texts will be discussed as separate stages in this and the next chapter, in many senses the stages 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 to answer questions about it and to aid in its transcription. If a text is collected, and saved for later computerization and annotation, the individual who collected it 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 as possible is recorded initially so that, at a later stage, those working with the text will be able to recover this information easily.

3.1 General considerations

As chapter 2 demonstrated, before the actual data for a corpus are 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. As was stressed at the start of chapter 2, however, it is important not to become too rigidly invested in the initial corpus design, since numerous obstacles and complications will be encountered while collecting data that may require changes in the initial corpus design: it may not be possible, for instance, to obtain recordings for all of 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 ICE Project provides a number of examples of logistical realities that forced changes in the initial corpus design. For instance, after the project began, it was discovered that not all of the regional groups involved in the project would be able to collect all of the text categories originally planned for inclusion in the corpus. However, rather than drop these regional groups from the overall project (or radically change the types of texts to be collected), it was decided that the groups should produce a shorter corpus and collect the categories that they were able to collect. Consequently, while a given regional component of the ICE Corpus may not contain telephone conversations, for instance, it will contain the required amount of some other part of the corpus (such as spoken dialogues) that can be compared with other regional components of the ICE Corpus containing comparable texts. Although this change means that the overall ICE Corpus will not be as complete as originally envisioned, it will nevertheless allow valid comparisons of the particular genres included in it.

Because collecting speech and writing involves quite different methodologies, the collection of each will be considered separately.

3.2 Collecting samples of speech

Speech is the primary mode of human communication. As a result, all kinds of speech have evolved: there are not only spontaneous multi-party 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 it, collecting data for the spoken part of a corpus is much more difficult and involved than collecting written samples. Consequently, there are numerous methodological considerations that must be addressed in the collection of speech.

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 turning on a tape recorder and having people converse. As anyone who has collected language data knows, as soon as you turn on a tape recorder and tell people to speak, you have to deal with the common scientific problem known as the "observer's paradox": by observing something, you change its natural behavior. And this is especially true with speech because many individuals are quite insecure about the way that they speak, and when they know that their speech is being monitored, they may adjust their speaking style and attempt to speak what they perceive to be "correct" English.

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 illegal in some locales and considered unethical by most linguists.[1] It is therefore imperative that anyone recording speech not only informs individuals that they are being recorded but obtains written permission from the individuals to use their speech in the corpus. This can be accomplished by having individuals sign a release form prior to being recorded.[2]

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: so little speech has been collected and included in corpora that it is not really possible to know empirically precisely what natural speech is really like. However, there are ways to increase the probability 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 they are participating in. 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 it is possible to study any kind of speech objectively, whether it is considered 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

[1] See Murray and Ross-Murray (1992 and 1996) for a discussion of the ethical and legal issues surrounding surreptitious recordings.

[2] With broadcast recordings, it is not usually necessary to obtain the permission of individuals to use their recordings but rather to contact the broadcast station, which owns rights to the program, and receive the station's written authorization to use the program. How to obtain copyright clearance for texts is discussed in section 3.3.

the most natural segment of speech from a much lengthier speech sample. How much speech needs to be recorded is determined by the type of speech being recorded and the length of segments that will be included in the corpus. If the goal is to create a corpus consisting of 2,000-word samples, for instance, then between ten and twenty minutes of speech is necessary to supply a single sample. The type of speech being collected will determine the amount of speech that is necessary. Spontaneous dialogues tend to consist of shorter speaker turns and many pauses and hesitations. Consequently, such conversations require a longer span of speech to reach the 2,000-word length requirement. On the other hand, monologues (especially scripted monologues) contain far fewer pauses and hesitations. Thus, less speech is needed to reach 2,000 words.

Work with ICE-USA has shown that it is most desirable for recordings of speech to be thirty minutes or longer. This length of conversation increases the probability that a natural and coherent 2,000-word speech sample can be extracted from this longer sample. The initial part of the conversation can 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.[3] Moreover, a lengthier segment allows the corpus compiler to select a more coherent and unified 2,000-word segment of speech to ultimately include in the corpus.

When it comes time to actually make recordings, whoever is making the recordings needs to follow a few basic principles to ensure the most natural recordings 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 far as possible, those making recordings should try to record individuals in the actual environments in which they typically speak. Samples included in ICE-USA include individuals speaking while eating a meal, working in offices, riding in a car, participating in class discussions, 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.

Because people are being recorded in their natural speaking environments, very often the quality of recordings is not as high as one would desire. If people are recorded while driving in a car, for instance, the recording will contain traffic noise. However, this noise can sometimes be edited out; in a recording of this type included in ICE-USA, the conversation itself was audible, even though there was considerable noise in the background. If high-quality recordings

[3] A similar conclusion was reached in recordings made for the British National Corpus. See Crowdy (1993: 261–2).

are desired, individuals can be recorded in a 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 details). However, in speech collected this way, while high-quality recordings will be obtained, the naturalness of the speech collected will be compromised. Therefore, it is probably wise to sacrifice recording quality in favor of natural speech.

One of the more elaborate and interesting ways of collecting speech was used to obtain speech samples for the British National Corpus (BNC).[4] 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). To keep track of the conversations they had, participants filled out a log book, 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.

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, so this is precisely the kind of speech one wants to gather. Likewise, with other types of speech, such as public speeches (especially if they are scripted), heavily edited speech is the norm.

In recording any kind of spoken English, the type of tape recorder to use is an important issue to consider. In the past, 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 widely available a variety of tape recorders that can make high-quality digital recordings. Either an analog or digital recorder will yield satisfactory recordings of multi-party dialogues: in pilot studies for the British National Corpus, for instance, Crowdy (1993: 261) reports that even though digital recorders produce high-quality recordings "under good recording conditions they were not significantly better than analogue recorders under the recording conditions likely to be encountered in this project [the British National Corpus]."

Despite this fact, however, there are distinct advantages to using digital rather than analog recorders. Analog cassettes age over time, and even if they are stored properly, they will eventually deteriorate in quality, rendering the recordings that were made unusable. Digital tapes, on the other hand, do not degrade as quickly. Although cassette tapes come in various lengths, they cannot record

[4] This method of collection was also used to collect samples for the Bergen Corpus of London Teenage English (COLT). See Haselrud and Stenström (1995) for details of this corpus.

more than sixty minutes of continuous speech on a single side (unless a special recorder is purchased that automatically begins recording on the second side of a cassette). If a fairly lengthy recording is being made, and a conversation has to be stopped to turn over the cassette, the naturalness of the recording can be seriously undermined. Digital cassettes, in contrast, can record up to four hours of continuous speech.

When working with spoken recordings, it is often desirable to work with digital recordings on a computer, for instance, to edit out unwanted background noise or to prepare the recording for inclusion on a CD for distribution as part of the corpus. Although the speech on an analog cassette can be digitized, a digital cassette can be transferred to disk directly with a minimal loss of recording quality. Presently, digital tape recorders and cassettes are more expensive than analog recorders and cassettes. However, current technology is going digital, and as it does, the cost of digital recorders and tapes will continue to decline. Therefore, unless cost is an important consideration, there is little reason for corpus compilers to use analog recorders anymore.

In addition to considering the type of tape recorder to use, it is equally important to consider the quality and type of microphone to make recordings with. An inexpensive 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, one that records an individual speaking directly into it. For larger groups, however, it is better to use omni-directional 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 that can be used 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 the telephone. High-quality microphones can be fairly expensive, but they are worth the investment.

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 ten recordings made, it may turn out that only four or five are usable. For instance, one recording made for ICE-USA contains two individuals who converse quite naturally for periods of ten minutes after which one of the individuals on more than one occasion remarks, "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. Since many televisions do not have audio output plugs, if the television is connected to a video cassette recorder or DVD player, a recording can be made by running a line from the audio output plug on the recorder or player.

In sending individuals out to make recordings, it is important to realize that making good recordings takes practice, and that as more recordings are made, the quality will increase as mastering the art of recording improves.

3.3 Collecting samples of writing

Although collecting samples of writing is considerably less complicated than collecting samples of speech, one additional and sometimes quite frustrating complication is encountered when collecting writing: copyright restrictions. Under the "fair use" provisions of current US copyright laws, it is acceptable to copy either all or part of a copyrighted text for purposes of non-profit academic research. However, fair use of copyrighted material is exceeded if a 2,000-word excerpt of a copyrighted article is computerized, for instance, and then distributed as part of a corpus, even if it will be used for non-profit academic research only. Therefore, the corpus compiler will need to get the written permission of any author whose writing is to be included in a corpus. This includes not just printed texts in magazines, journals, and books, which will always be marked as copyrighted, but letters, online articles, and other unpublished works as well, which are automatically assigned copyright. In fact, the only written materials in the United States that are not copyrighted are any US government publications and works that are beyond seventy years from the author's date of death. Although many countries have signed international copyright agreements, there is variation in practice. Therefore, the corpus compiler needs to consider not just the laws of the country in which a corpus is being assembled but the laws of the country in which the written documents were created and copyrighted.

When making recordings of speech, one obtains permission from speakers before they are recorded. Therefore, there is no time wasted making recordings and then finding out that the speakers will not give permission for their recordings to be used in the corpus being created. However, with written texts, this luxury is not available: 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 or ICE) a letter might be mailed requesting permission but no reply to the letter is ever received. In cases like this, current copyright law allows for the use of such materials, provided that it can be established that a reasonable effort was made to obtain permission. But a cautious corpus compiler might wish to avoid any future legal problems from authors of texts from whom no written permission was obtained. For this reason, no written text was included in ICE-USA without the written permission of its author (or authors).

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 has 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 for non-profit academic research only, this should be clearly stated in any letter of inquiry requesting permission to use copyrighted material and many publishers will 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 some cases, obtaining permission to use copyrighted material has proven so onerous that some corpora can only be used on-site where they were created. For instance, because permission could not be obtained to distribute the English–Norwegian Parallel Corpus to any interested researcher, it is only available to researchers able to travel to the University of Oslo, where it was created and is now housed. Ask anyone who has created a corpus containing copyrighted material, and they will state that their greatest frustration in compiling the corpus was dealing with copyright complications.

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. Because whatever writing collected

will ultimately need to be computerized, a good place to start is journals and magazines published on the World Wide Web: an increasing number of periodicals are now available online, and once permission is obtained to use an article in a periodical, the article can be easily downloaded in computerized form.

Of course, including texts of this type in a corpus raises a methodological issue: are electronic texts essentially the same as traditionally published written texts? With newspapers, this is less of an issue, since many newspapers publish online versions of articles that are identical to those included in printed editions.[5] But many written texts exist exclusively online, and if a corpus contains, for instance, five electronic texts in the genre of scientific writing and five traditionally published texts in the same genre, do we have one genre here or two subgenres: electronically published scientific writing and traditionally published scientific writing? And since anyone can publish an article on the World Wide Web, is an electronic text published on an individual's personal home page any different from a text that has gone through the editorial process and been published in a journal, magazine, or book? These are important questions, and in answering them, some might conclude that electronically published texts are different enough from traditionally published works that the two types of texts should not be mixed together in a single genre within a corpus. But there is no denying that the trend towards electronic publishing is increasing, and even though traditional periodicals such as journals and magazines will probably never disappear, including at least some electronically published writing in a synchronic corpus accurately reflects the current state of writing. A further advantage of electronically published writing is that it allows the corpus compiler to avoid the tedious job of computerizing printed texts, either by typing them in by hand or by scanning them with optical scanners, which ease the process of computerizing written texts but which still produce an electronic copy of the text albeit with a number of scanning errors which have to be manually corrected.

If a corpus compiler decides to include electronic texts in a corpus, a good way to find them is to use an indexed search engine such as Yahoo (http://www.yahoo.com), that is, a search engine that organizes information on the World Wide Web by putting it into categories. The opening page of Yahoo is divided into a series of subject areas, such as news media, science (i.e. natural science), social science, and arts and humanities – all genres commonly found in corpora. Selecting "science" yields a number of choices, including "journals," which in turn provides links to numerous online journals in areas such as zoology (http://dir.yahoo.com/Science/Biology/Zoology/Journals/) and biology (http://dir.yahoo.com/Science/Biology/Journals/). Each of the journals offers a number of potential texts for inclusion in a corpus.

Alternatively, a search engine such as AltaVista (http://www.altavista.com/) that searches for strings on the World Wide Web can be used. A search on

[5] Even though most newspapers publish both printed and online versions of each issue, it is worth checking the original printed edition with the online edition just to make sure that the two versions match.

AltaVista of the key words "fiction" and "magazine" produced thousands of sites, and among these sites were many fiction magazines containing full-text pieces of fiction. So many online texts exist that it is now possible to create an entirely online corpus; that is, a corpus consisting exclusively of links to texts that could be subjected to the same kinds of linguistic analyses now possible only with corpora made available on disk (cf. Davies 2001 for an example).

Even though a corpus compiler may not want to include purely electronic publications in a corpus, there are a number of traditionally printed texts that are also available in electronic form. For instance, some newspapers have released CD-ROMs containing previously published articles. It is also increasingly common, as was noted above, to find newspapers with websites on which they publish all or some of the articles in their daily editions. If the paper does not have its own website, very often its articles will be in the LEXIS/NEXUS database, which many academic libraries subscribe to and from which the articles in many newspapers (and other written texts too) can be downloaded.

There are various ways to obtain printed texts. For instance, learned writing in the areas of the humanities, natural sciences, social sciences, and technology can be collected by randomly selecting texts from the current periodicals sections of libraries. Other types of texts can be obtained in this manner as well. Various periodical indexes and online catalogues can also be consulted for references to books and articles. When gathering printed texts, it is important that as good a photocopy as possible of each text is made so that if the text is eventually included in the corpus, it can be scanned as accurately as possible: poor photocopies tend to produce more numerous scanning errors.

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 might be 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 name given to a text sample is short and mnemonic and gives 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 tape recording to work with and has to 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 future users of the corpus will want to investigate. Because ICE-USA is a multi-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 (Powell and Simpson 2001: 35).

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

British National Corpus and ICE, do not contain information on many speakers and writers.

After all of 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 is kept about each text to be considered for inclusion in the corpus.

3.5 Computerizing data

Computerizing spoken and written texts for inclusion in a corpus is a very labor-intensive part of creating a corpus. Transcribing speech is an extremely lengthy process, requiring the transcriber to listen to the same segments of speech over and over again until an accurate transcription is achieved. Even though printed texts can be scanned into a computerized format, there will always be scanning errors – so many errors, in fact, that some corpus compilers have abandoned scanners altogether and have found that retyping texts is a quicker and more efficient way to computerize them, particularly if it is not possible to obtain a clear printed copy of the text to be scanned (which is essential for accurate scanning). For both spoken and written texts, proofreading a large corpus of texts after they have been computerized is painstaking work, and no matter how carefully it is done, errors will remain.

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 a word-processing 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, it is best from the onset to save texts in ASCII (or text) format, since this is the standard format used for texts included in corpora, and to use a simple text editor to work with texts rather than a standard word-processing program: even though such programs claim to create pure ASCII files, very often they insert formatting commands that can be problematic when a given corpus is analyzed by a tagging or concordancing program.

The ASCII file format has both advantages and disadvantages. The main advantage is that ASCII is a universally recognized text format, one that can be used with any word-processing program and the numerous software programs designed to work on corpora, such as taggers and parsers (cf. sections 4.3 and 4.5) and concordancers (cf. section 5.3.2). The disadvantage is 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 create 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 Old English ash "æ" (Kytö 1996). This system of symbols is specific to the Helsinki Corpus. There exists a successor to ASCII, Unicode, containing an expanded character set able to represent all of the characters found in the languages of the world. Unfortunately, this standard has not been around long enough to have replaced ASCII in the corpus linguistics community. As a result, most corpus compilers use some kind of system of annotation to represent non-ASCII characters (cf. section 4.1), though with the emergence of the Text Encoding Initiative (TEI), this can be done more systematically so that each corpus project does not have to create a unique system of annotation (different from somebody else's system) to represent non-ASCII characters. The TEI has developed a formalism, called a "writing system declaration," for identifying the specific character set used in a given electronic document (see "Overall Structure of Writing System Declaration" in Sperberg-McQueen and Burnard 1994a, http://etext.lib.virginia.edu/bin/tei-tocs-p3?div=DIV2&id=WDOV). Whichever system is used, because information will be lost when a text is encoded in ASCII format, it is important that a printed copy of every written text is kept on file, so that when the text is annotated, the appropriate markup can be added to it to mark non-ASCII characters.

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 edn.) allows users to work with an entire corpus saved in a single file. But organizing a corpus into a series of directories and subdirectories 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 ICE-USA.

ICE-USA consists of 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 subdirectories 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.

The remainder of figure 3.1 contains additional subdirectories for the specific types of dialogues that were gathered (for reasons of space, the other subdirectories are not included): business transactions, classroom discussions, political debates, and so forth. And within each of these directories would be the individual texts collected to fill the category. Since an individual text goes through many stages of analysis, various versions of a text are kept in separate directories. Texts that are in the process of being transcribed or scanned are kept in the "draft" directory. Once a draft version of a text has been fully annotated with

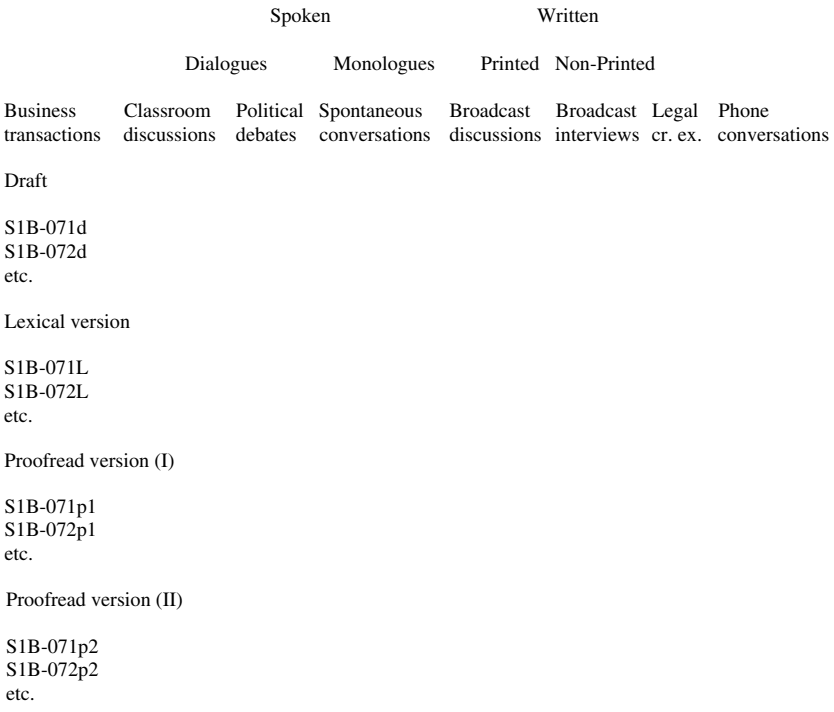


Figure 3.1 Partial directory structure for American component of ICE

"structural" markup (cf. section 4.1), it is moved into the "lexical" directory to indicate that the text is ready for inclusion in a lexical version of ICE-USA (i.e. a version of ICE-USA containing the text and structural annotation). The text then undergoes the first round of proofreading after which it is placed in the directory "Proofread version (I)." A second round of proofreading is done after completion of the entire corpus so that any editorial changes made during the creation of the corpus can be incorporated into the final version of the corpus, which is placed in the directory "Proofread version (II)."

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 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 over and over again 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.

3.6 Computerizing speech

Traditionally, speech has been transcribed using a special transcription machine which has a foot pedal that stops and starts a 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 be able to replay segments automatically, and anyone attempting to transcribe a tape using a traditional tape recorder is wasting their time and adding unnecessary hours to their transcription efforts.

Because of recent advances in computer technology, it is now possible to use software programs designed specifically to transcribe speech that has been digitized. "VoiceWalker 2.0" was developed to aid in the transcription of texts included within the Santa Barbara Corpus of Spoken American English (http://www.linguistics.ucsb.edu/resources/computing/download/download.htm). "SoundScriber" is a similar program used to transcribe texts that are part of the Michigan Corpus of Academic Spoken English (MICASE) (http://www.lsa.umich.edu/eli/micase/soundscriber.html). 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. The advantage of software like this is that it has made obsolete the need for cassette transcription machines, which because of all of their moving parts tend to break down after frequent use. The disadvantage of working with digitized speech is that high-quality digital recordings can require very large amounts of disk space for storage. But this

problem can be minimized if recordings are saved at settings resulting in lower quality (but still highly audible) recordings. ICE-USA saves digital recordings as .wav files in Windows PCM format, mono, 16 bit, 16000 Hz. A typical 2,000-word sample of speech saved with these specifications will require 20–30 MB of hard disk space. If the same file were saved in stereo at CD quality, it would require 80 MB (or more) of hard disk space. Since the ICE Project began, a new file format, .mp3, has gained popularity, primarily because it creates audio files much smaller than .wav files: a 26 MB ICE .wav file, for instance, was converted into a 2.5 MB .mp3 file, and there was very little loss of sound quality.

To digitize analog cassettes, all that is needed is a sound board and a program that can digitize speech, such as Syntrillium's "Cool Edit" (available as shareware at http://www.syntrillium.com/cooledit/index.html). To use a program such as "Cool Edit" to digitize speech, one simply connects the audio output of the cassette recorder to the audio input of the sound board and has the program digitize the speech sample as it is playing on the cassette player. After the sample is digitized, it can be saved on either a hard drive or some other media, such as a CD.

While transcription machines and software can ease the process of transcribing speech, there is no getting around the fact that speech has to be manually transcribed. Although 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" (http://www.voice-recognition.com/1998/products/dragon/dictate.html), 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 recognize continuous speech. There now exist a number of programs, such as "Dragon NaturallySpeaking" (http://voicerecognition.com/1998/products/dragon/vrstandard.html), that have large lexicons and, with training, can quite accurately recognize carefully articulated monologues. Current programs under development have progressed to the point where they are beginning to work with certain kinds of dialogues. Nguyen et al. (1999) describe a system that recognizes the kinds of monologues and dialogues spoken on broadcast news programs. Built into the system is the ability to adjust to the speech of males and females and to handle the various "dysfluencies" that occur in spontaneous speech: repetitions, pauses, partially articulated words, and vocalized pauses such as "uh" and "uhm." Although speech-recognition programs may never be able to take a multi-party dialogue, for instance, and in a matter of seconds produce an accurate transcription of the dialogue, it is likely that at some time in the future the technology may help ease the time-consuming task

of transcribing speech manually. And even if such systems are not completely accurate, they could at least provide a draft transcription of a conversation that could then be edited.

In terms of transcription time, speech runs a continuum, with multi-party dialogues with numerous overlaps taking the most time to transcribe, and scripted monologues taking the least time to transcribe. For ICE-USA, a 2,000-word multi-party dialogue takes on average fifteen to twenty hours to transcribe, annotate with markup, and proofread. A scripted monologue, on the other hand, takes much less time (approximately five to eight hours). The remaining kinds of speech, such as broadcast dialogues or telephone conversations, require transcription times somewhere between these two extremes.

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. In creating the Corpus of Spoken Professional English, Barlow ("CPSA Description": http://www.athel.com/corpdes.html) made a number of compromises. This is not a corpus of speech that Barlow recorded and transcribed. Instead, it consists entirely of transcriptions of academic meetings and White House press conferences that are in the public domain and that were created by a third party. Consequently, while the transcriptions, according to Barlow, appear "relatively unedited and thus include hesitations, false starts, and so on," ultimately there is no way to determine how faithful they are to the original speech. Given the cost and effort of creating a corpus of speech, it is understandable why corpora of this type exist, and undoubtedly, they will yield much valuable information about speech. But there will always be an element of doubt concerning results taken from a corpus created in this manner.

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 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 amount of time it takes to annotate a text with information about intonation (cf. section 4.1). The advantage is that a tremendous amount of 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 of 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 annotation to mark 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. annotation marking 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 far 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 discussion.

3.6.1 Representing speech in standard orthographic form

Many of the expressions found in speech can easily be represented within the constraints of standard orthography, particularly if annotation is added to a transcription to mark the particular feature being transcribed.

3.6.1.1 Vocalized pauses and other lexicalized expressions

Speech contains a group of one- or two-syllable utterances sometimes referred to as "vocalized pauses," verbal expressions that allow the speaker to pause and plan what is to be said next. These expressions form a closed class and include pronunciations such as [ə] or [a:], [əm], and [əhə]. Although there is no universally agreed spelling for these expressions, in the ICE corpora, these pronunciations are transcribed, respectively, as *uh, uhm,* and *uhuh*.

Speech also contains a group of expressions that can be used to answer *yes/no* questions or to express various kinds of emotions. To represent these expressions in a transcription, the Michigan Corpus of Academic Spoken English (MICASE) has developed a number of orthographic representations. For instance, for the various expressions that can be used to answer *yes/no* questions in English, MICASE has created the following orthographic representations:

yes: *mhm*, *mm*, *okey-doke*, *okey-dokey*, *uhuh*, *yeah*, *yep*, *yuhuh*
no: *uh'uh*, *huh'uh*, *'m'm*, *huh'uh* ("MICASE Transcription and Spelling Conventions" http://www.lsa.umich.edu/eli/micase/transcription.html)

To represent various kinds of verbal expressions that people use, MICASE has representations such as *ach, ah, ahah, gee,* and *jeez.*

3.6.1.2 Linked expressions

In speech, there are examples of expressions that are spelled as two separate words, but pronounced as one word. In fictional dialogue, they are sometimes spelled as *gotta* (for *got to*), *hafta* (for *have to*), and *gonna* (for *going to*). If an expression such as *gotta* is viewed merely as a phonetic merging of *got* and *to*, then all instances of this pronunciation can simply be transcribed as *got to*. However, if *gotta* is felt to be a single lexical item, then it should be transcribed differently than *got to*. How these words are transcribed depends crucially upon which of these two analyses one accepts. If, for instance, all instances of *gotta* and *got to* are transcribed as *got to*, then whoever uses the corpus will not be able to recover the distinct forms of this expression. For this reason, the ICE project includes *gotta, hafta,* and *gonna* in its transcriptions of spoken texts. Other corpus projects, such as MICASE, include additional examples, such as *kinda, sorta,* and *lotsa* ("MICASE Transcription and Spelling Conventions": http://www.lsa.umich.edu/eli/micase/transcription.html).

3.6.1.3 Partially uttered words and repetitions

Speech (especially unscripted speech) contains a number of 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 s1a-040-050)

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, <.> and </.>, that labels the expression as an instance of an incomplete word. Other corpora, such as the London–Lund Corpus, sometimes provide a phonetic transcription of such utterances.

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 (cf. section 5.3.2), the text analysis program used in the ICE Project.

3.6.1.4 Unintelligible speech

Very often when people speak, their speech is not intelligible. 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 any transcription of spoken dialogues will therefore encounter instances where speech cannot be transcribed because it is not understandable. In the ICE project, stretches of speech of this type are annotated as an "uncertain transcription" or as an "unclear word or syllable." In the example below, the tags <?> and </?> surround the word *them* because the transcriber was uncertain whether this was actually the word the speaker uttered.

<$C> <#9:1:C> <sent> What was Moses doing going off in <?> them</?> jeans

3.6.1.5 Punctuation

Because speech is segmented by intonation rather than by punctuation, if speech is transcribed without any punctuation, it can be very difficult to read, as sentences and utterances are run together and difficult to separate:

<$A> you can fish off our dock interesting not this summer last summer I was sitting in the boat listening to the radio and a man walked to the end of the peer put his tackle box down put some I think artificial bait on he started casting and he caught two bass two peeper bass within while I was sitting there watching him for half an hour it depends on on the condition you don't have to go a long ways in the evening they would be in the shallow water <unintelligible> (ICE-USA)

To make speech easier to read, it is tempting to add standard punctuation to a spoken transcription. However, by punctuating speech, the corpus compiler is in a sense "interpreting" the spoken text for future users of the corpus and therefore making decisions that the users really ought to make themselves as they analyze a spoken text and (if possible) listen to a recording of it. For this reason, the ICE project includes only a very limited number of punctuation marks: apostrophes for contractions and possessives, capitalization of proper

nouns and words beginning text units, and hyphens for hyphenated words. If the corpus compiler wishes to segment spoken texts in transcriptions, it is far better to include prosodic transcription in a corpus; that is, annotation that marks such features of intonation as pauses, tone unit boundaries, and pitch changes (cf. section 4.1).

3.6.1.6 Background noise

As speakers converse, they very frequently laugh, cough, sneeze, and make other non-linguistic sounds. In many transcription systems, these noises are simply ignored and not transcribed. In the ICE project, such sounds are enclosed in markup: <O>cough<O/> or <O>sneeze<O/>. In other corpora, comparable conventions are used.

3.6.1.7 Changing the names of individuals referred to in spoken texts

In any conversation, speakers will address each other 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 are changed in transcriptions to protect the privacy of the individuals conversing and the people they are conversing about. And if the recordings accompanying the transcriptions are to be made available as well, any references to people's names 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 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).

3.6.2 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 ASCII 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 *yeah* 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. However, instead of representing these overlaps with markup (as an SGML-conformant system would do; cf. section 4.1), this system attempts to indicate them iconically by vertically aligning the parts of the conversation that overlap to give the reader 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 should be represented in tabular form, a manner of presentation that provides both an accurate and iconic representation of a conversation (cf. also Meyer, Blachman, and Morris 1994). 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 already did figure three
we did it before	

(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. Speakers S1, S2, and Sy's 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 inconicity 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 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 idiosyncrasies 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 the markup that they contained. Nevertheless, one of the great challenges that corpus linguists must face is the development of software with user interfaces that permit users to browse spoken corpora easily and effortlessly.

3.7 Computerizing written texts

Because written texts are primarily linear in structure, they can easily be encoded in an ASCII text format: most features of standard orthography, such as punctuation, can be maintained, and those features that cannot, such as font changes or special characters, can be annotated with an SGML-conformant tag. Although computerizing a written text takes time, particularly if it has to be scanned, this process is far less time-consuming than computerizing a spoken text. However, if one is creating a historical corpus and thus working with texts from earlier periods of English, converting a text into electronic form 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 have to 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*, Markus (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: eleven versions in English, four in Latin, and two in French (Markus 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 which 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 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 eighty-eight versions of Chaucer's *Canterbury Tales* and allows various kinds of comparisons between the differing versions (Robinson 1998). 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.[6] 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 is ideal, but when working with texts from earlier periods, the corpus compiler has to make compromises given the kinds of texts that exist.

Fortunately with modern-day texts, none of these complications exists, and the process of computerizing a written text involves mainly getting the text into electronic form. As was noted in section 3.3, if a written text is not available in electronic form, a printed copy of the text can be converted into an electronic format with an optical scanner. There are two variables that will affect the success that one has with an optical scanner. The first variable is the quality of the original printed text. If the text is blurred or does not contain a distinct typeface, when scanned, it will contain numerous typographical errors that may take longer to correct than it would to simply retype the entire text by hand. The second variable is the quality of scanner and the OCR (optical character recognition) software used with it.

There are two types of scanners: form-feed scanners and flatbed scanners. Experience with ICE-USA has shown that flatbed scanners are slightly more accurate than form-feed scanners but that satisfactory results can be obtained with the use of a relatively inexpensive form-feed scanner if the printed copy to be scanned is of high quality and no attempt is made to scan pages with columns, a common format in newspapers and magazines and a format that cannot be handled by the OCR software commonly bundled with inexpensive form-feed scanners. However, this problem can be solved if columns are copied, clipped, and scanned separately.

Because so many kinds of written texts are now available on the World Wide Web, it makes more sense to obtain as many written texts as possible from relevant websites. However, even though such texts can be easily downloaded in electronic form, in many cases they will contain as much HTML coding

[6] Because edited versions of older texts are always copyrighted, the corpus compiler will need to obtain copyright clearance for use of the version. This can greatly complicate the creation of a historical corpus. ARCHER, for instance, has not been publicly released because copyright clearance could not be obtained for many of the texts included in the corpus.

as text. To attempt to manually delete this coding will require a considerable amount of time and effort. Fortunately, there exists software (e.g. "HTMASC" http://www.bitenbyte.com/) that can automatically strip HTML coding from texts and produce in seconds an ASCII text file with no coding.

3.8 Conclusions

The process of collecting and computerizing texts is, as this chapter has demonstrated, a labor-intensive effort. Collecting speech requires the corpus creator to become, in a sense, a field linguist and go out and record individuals in the various locations where speech takes place: homes, offices, schools, and so forth. And once speech is recorded, considerable time and effort must be expended transcribing it. Written texts are somewhat easier to collect and encode in electronic format, but obtaining copyright clearance for copyrighted texts can be a very complicated undertaking; if texts from earlier periods are being collected and computerized, decisions will need to be made about what exactly goes into the version of the text that is to be included in a corpus if, for instance, more than one manuscript version of the text exists. Once the process of collecting and computerizing texts is completed, texts to be included in a corpus are ready for the final stage of preparation – annotation – a topic discussed in the next chapter.

Study questions

1. What makes collecting "natural" speech so difficult?
2. To record speech, those creating the British National Corpus gave individuals tape recorders and had them record conversations they had with others. Why is this potentially a better way of collecting speech than sending out a research assistant with a tape recorder and having him or her make the recordings?
3. What are some of the advantages and disadvantages of including in a corpus written texts obtained from the World Wide Web?
4. Why, when collecting texts for a corpus, is it important to keep records documenting ethnographic information for writers and speakers whose texts will be included in the corpus?
5. What problems do texts taken from earlier periods of English pose when it comes time to computerize the texts?

4 Annotating a corpus

For a corpus to be fully useful to potential users, it needs to be annotated. There are three types of annotation, or "markup," that can be inserted in a corpus: "structural" markup, "part-of-speech" markup, and "grammatical" markup.[1]

Structural markup provides descriptive information about the texts. For instance, general information about a text can be included in a "file header," which is placed at the start of a text, and can contain such information as a complete bibliographic citation for a written text, or ethnographic information about the participants (e.g. their age and gender) in a spoken dialogue. Within the actual spoken and written texts themselves, additional structural markup can be included to indicate, for instance, paragraph boundaries in written texts or overlapping segments of speech in spoken texts. Part-of-speech markup is inserted by a software program called a "tagger" that automatically assigns a part-of-speech designation (e.g. noun, verb) to every word in a corpus. Grammatical markup is inserted by a software program called a "parser" that assigns labels to grammatical structures beyond the level of the word (e.g. phrases, clauses).

This chapter focuses on the process of annotating texts with these three kinds of markup. The first section discusses why it is necessary for corpora to be annotated with structural markup and then provides an overview of the various systems of structural markup that have evolved over the years as well as the various tools that have been developed for inserting markup in corpora. The next two sections are concerned with the process of tagging and parsing a corpus. These sections will describe some of the more common taggers and parsers that are available, provide descriptions of what tagged and parsed texts look like, detail how the various taggers and parsers available reflect different views of English grammar, and discuss some of the constructions that lead to tagging and parsing errors. In addition, other types of tagging, such as semantic and discourse tagging, are briefly described. Even though much of the underlying work on taggers and parsers has been done by computational linguists doing research in the area of natural language processing, this chapter will be primarily concerned with how tagged and parsed corpora can be of benefit to corpus linguists interested in creating and analyzing corpora for purposes of descriptive

[1] There is also annotation, such as semantic annotation, that can be used to mark higher-level structures larger than the word, clause, or sentence. Cf. section 4.4.3 for details.

linguistic research; theoretical issues of computation will be broached only as they clarify issues related to linguistic theory and description.

4.1 Structural markup

A document created by any word-processing program can be formatted in numerous ways: font types and sizes can be varied, sections of the text can be italicized or placed in boldface, footnote indicators can be displayed in superscript form, and so forth. However, as soon as this document is converted into ASCII (or text) format, all of this formatting is lost, and only pure text can be displayed (cf. section 3.5 for a discussion of ASCII text format). Therefore, for texts passed from computer to computer in ASCII format, some system of markup needs to exist so that anyone receiving a text can "reconstruct" how it originally looked. And in language corpora, this issue becomes especially important: not only are there features of written corpora that will need to be preserved, but there are characteristics of speech (such as overlapping speech) that need to be marked so that they can be easily identified by the user of the corpus.

In earlier corpora, a fairly minimal and non-standardized set of markup was used to annotate various features of speech and writing. For instance, in the London–Lund Corpus, speakers were identified with alphabetic letters (A, B, C, etc.) followed by colons, and overlapping segments of speech were enclosed with asterisks:

A: yes that I think you told me *I*
B: *and* none of them have been what you might call very successful in this world
(LLC S.1.13 98–100)

More recently, as electronic documents have proliferated, a standard for the markup of electronic documents has developed. This standard, known as Standard Generalized Markup Language (SGML), is not a series of predetermined symbols, or "tags," as in the example above, but instead a "metalanguage" that provides a mechanism for describing the structure of electronic documents. For corpus compilers, the main advantage of placing SGML-conformant markup in their corpora is that information about a corpus can be consistently and unambiguously described and maintained as the corpus is transferred from computer to computer. Although the idea behind SGML is simple and straightforward, the actual implementation of an SGML-based system of markup can be quite complex. For this reason, this section will provide only a general overview of SGML.

SGML markup is purely "descriptive," that is, it does not tell the computer to perform any operations on a text but instead consists of "markup codes which simply provide names to categorize parts of a document" (Sperberg-McQueen and Burnard 1994b: http://etext.lib.virginia.edu/bin/tei-tocs-p3?div=DIV1&id=SG). The example below shows how the previous excerpt would be annotated with the SGML-conformant "structural tags" used in the ICE project

(cf. Nelson 1996 for a complete discussion of the structural markup used to annotate ICE texts).

<$A><#:1> yes that I think you told me <{><[>I<[/>
<$B><#:2> <[>and<[/><{/> none of them have been what you might call very successful in this world

Some of the markup provides a description of the structure of the conversation. Two of the tags, <$A> and <$B>, indicate who is speaking at a given time in the conversation; two other tags, <#:1> and ><#:2>, set off and number consecutively what are called "text units", that is, segments of a spoken or written text that are either grammatical sentences or, in speech, utterances: stretches of speech that may be grammatically incomplete but that form a coherent grammatical unit. For instance, the expression *Yeah* is not a grammatical sentence but an utterance: when used in a conversation as a response to a *yes/no* question, it makes perfect sense and is semantically coherent. Other tags in the excerpt describe individual features of the conversation. Two tags, <{_> and <{/>, mark the beginning and end of a speech segment containing overlapping speech; within this lengthier segment are two sets of additional tags, <[_> and <[/>, that mark the two individual segments of speech in each speaker turn that overlap. The ICE Project has developed markup for written texts as well to annotate such features of writing as italics, boldface, small caps, superscript and subscript symbols, and foreign letters not represented in ASCII format.

Typically, SGML-conformant tags contain a start tag and an end tag, as was the case in the above excerpt with the tags used to mark overlapping speech. However, sometimes only one tag is necessary. For instance, a speaker identification tag simultaneously marks the beginning of a new speaker's turn and the end of a previous speaker's turn. Likewise, a text unit tag marks the beginning of a new text unit and the end of a previous one.

In addition to containing markup placed directly in a text, ICE texts also begin with file headers, that is, a series of statements enclosed in ICE tags that provide, for instance, bibliographic information about written texts, or ethnographic information about individuals in spoken texts. Below is the beginning section of a file header for a written text included in the British component of ICE:

```
<text.info>
        <file.description>
                <textcode>ICE-GB-W2A-008</textcode>
                <number.of.subtexts>1</number.of.subtexts>
                <category>printed;informational:popular:humanities</category>
                <wordcount>2101</wordcount>
                <version>tagged by TOSCA tagger using the ICE tagset</version>
                <free.comments>                </free.comments>
        </file.description>
.
.
.
<text.info>                                                    (Nelson 1996: 51)
```

The tags in the above header provide important descriptive information about the text, such as the genre in which the text has been classified, the number of words that the text contains, and a statement that the TOSCA tagger assigned part-of-speech designations taken from the ICE tagset (cf. section 4.2). It must be emphasized that the tags used in the ICE Corpus are not SMGL tags per se, but rather SGML-conformant tags: the names and symbols used in the tags above were developed by the ICE Project within the constraints stipulated by SGML.

The tags for the ICE Project were developed in the late 1980s and therefore do not reflect recent work done by the Text Encoding Initiative (TEI) to develop a comprehensive SGML-conformant markup system for electronic documents, particularly those used in the humanities and in language corpora. As Burnard (1995) notes, an electronic document annotated with TEI-conformant markup would contain three types of "tagsets": "core tagsets" (a series of tags, such as those associated with file headers or paragraph divisions, available for insertion in any document); "base tagsets" (a series of tags associated with particular kinds of texts, such as verse, drama, or transcribed speech); and "additional tagsets" (any tags the user wishes to add to the core and base tagsets already in a document). The advantage of the TEI system is that it provides both a standardized set of tags for insertion in a document and the flexibility for the insertion of tags designed by the corpus compiler. Since it appears that TEI is the emerging standard for the markup of corpora (the entire British National Corpus was annotated with TEI markup), anyone contemplating the creation of a corpus ought to give serious consideration to using the TEI standards for the markup of texts. The more standardized creators of corpora can be in the markup they insert, the easier it will be for users to analyze corpora. A full description of the TEI system can be found in Sperberg-McQueen and Burnard (1994a).

More recently, there has been interest in the corpus linguistics community in a newly emerging markup system: Extensible Markup Language (XML). XML is a restricted version of SGML that has been designed mainly for use in web documents. The American National Corpus (a corpus of American English comparable to the British National Corpus in length and coverage) plans to use XML to annotate texts included in the American National Corpus ("American National Corpus": http://www.cs.vassar.edu/~ide/anc/). The EAGLES Project has an extensive working document to develop a "Corpus Encoding Standard" that describes how to use XML to annotate a corpus ("XCES: Corpus Encoding Standard for XML": http://www.cs.vassar.edu/XCES/). And the TEI Project has been working to incorporate XML within its standards. At this point, it is not clear how widely used XML will become in the corpus linguistics community. But a clear advantage of XML is its ability to be used on web pages, and as the web develops, it will become an increasingly important medium for the distribution of corpora.

There is one additional type of annotation that can be incorporated within an SGML-conformant system of markup but that has typically not been: annotation

that is used to mark features of intonation in spoken texts, such as changes in pitch or tone unit boundaries. The example below illustrates the system of annotation used in the Santa Barbara Corpus of Spoken American English (for additional details, cf. Chafe, Du Bois, and Thompson 1991: 75; Du Bois, Schuetze-Coburn, Cumming, and Paolino 1993):

G: For ^most people, it's ^celebration,\
for ^me =,_
it's .. it's a% ^ti=me,_
to=__
(H) to 'get in ^be=d,\/ (Chafe, Du Bois, and Thompson 1991: 77)

In the above example, all of the annotation indicates particular characteristics of intonation, such as falling pitch (\), primary accent (^), lengthened syllables (=), and a short pause (..). While prosodic annotation provides a more accurate rendering of the flow of speech than punctuation does (cf. section 3.6.1.5), inserting the necessary prosodic markup can increase transcription time significantly – in some cases, doubling or tripling the time of transcription. To minimize transcription time, the ICE project decided to mark two lengths of pause only – short and long – and to dispense with any other prosodic markup.

Although SGML-conformant systems of annotation provide important descriptive information about a text, they can pose numerous difficulties for both corpus compilers and corpus users. First of all, because much markup has to be manually inserted, if a corpus compiler wants to exploit the full possibilities of the TEI system, for instance, he or she will need considerable resources to hire assistants to insert the markup. There are tools that can assist in the insertion of markup. To assist in the annotation of ICE texts, the ICE Markup Assistant was developed (Quinn and Porter 1996: 65–7). This program uses a series of WordPerfect macros to insert ICE markup in texts. These macros are inserted either manually or automatically, depending upon how easy it is to predict where a given tag should be inserted. For instance, because overlapping segments of speech occur randomly throughout a conversation, their occurrence cannot be predicted and they therefore have to be inserted manually. Conversely, in a written text, text unit tags can be inserted automatically at sentence boundaries, and then those that are erroneously marked can be post-edited manually. A range of SGML resources can be found on the W3C website: http://www.w3.org/MarkUp/SGML/.

Another way to minimize the amount of time it takes to annotate texts is to use a reduced system of annotation. The ICE Project reduced the amount of structural markup required in ICE texts to the most "essential" markup for those ICE teams lacking the resources to insert all of the ICE markup that had been developed (Meyer 1997). Likewise, there exists a version of TEI called "TEI Lite," which contains a minimal set of TEI conformant markup (Burnard and Sperberg-McQueen 1995). In determining how much markup should be included in a text, it is useful to adopt Burnard's (1998) "Chicago

pizza" metaphor. One can view markup as toppings on a pizza: the particular "toppings" (i.e. markup) that are added depend upon both what the corpus compiler sees as important to annotate and what resources are available to insert this annotation. Although some object to this view – Cook (1995), for instance, views annotation used in spoken texts not as an embellishment, an extra topping, but crucial to the interpretation of a speech event – realistically the corpus compiler has to draw a line in terms of how detailed texts should be annotated.

While texts annotated with structural markup greatly facilitate the automatic analysis of a corpus, the user wishing to browse through a text (particularly a spoken text) will find it virtually unreadable: the text will be lost among the markup. The British component of ICE (ICE-GB) works around this problem by enabling the user to select how much markup he or she wishes to see in a text: all, none, or only some. Likewise, some concordancing programs, such as MonoConc Pro 2.0 (cf. section 5.3.2), can display markup or turn it off. One of the future challenges in corpus linguistics is the development of tools that provide an easy interface to corpora containing significant amounts of annotation. The introduction of XML markup into corpora offers potential help in this area, since as web browsers are designed to read XML annotated documents, they will be able to convert a corpus into the kinds of documents displayed by browsers on websites.[2]

The discussion in this section might lead one to think that there has been an endless proliferation of markup systems in corpus linguistics to annotate corpora with structural markup. However, it is important to realize that all of these systems have one important similarity: they are all SGML-conformant. The ICE system was developed to annotate ICE documents, the TEI system a wider range of spoken and written corpora as well as various kinds of documents commonly found in the humanities, and XML various corpora for possible distribution on the World Wide Web. Thus, all of these markup conventions are simply instantiations of SGML put to differing uses.

4.2 Tagging a corpus

In discussing the process of tagging a corpus, it is important, first of all, to distinguish a "tagset" – a group of symbols representing various parts of speech – from a "tagger" – a software program that inserts the particular tags making up a tagset. This distinction is important because tagsets differ in the number and types of tags that they contain, and some taggers can insert more than one type of tagset. In the example below, the Brill tagger (described in Brill 1992), adapted for use in the AMALGAM Tagging Project, has been used

[2] Cf. Edwards (1993) for further information on how markup can be made more readable, particularly markup used to annotate spoken texts.

to assign part-of-speech designations to each word in the sentence *I'm doing the work*:[3]

I/PRON(pers,sing)
'm/V(cop,pres,encl)
doing/V(montr,ingp)
the/ART(def)
work/N(com,sing)
./PUNC(per)

In this example, which contains tags from the ICE tagset (Greenbaum and Ni 1996), each word is assigned to a major word class: *I* to the class of pronouns, *'m* and *doing* to the class of verbs, *the* to the class of articles; *work* to the class of nouns; and the end stop to the class of punctuation. In parentheses following each major word class are designations providing more specific information about the word: *I* is a personal pronoun that is singular; *'m* is the enclictic (i.e. contracted) present-tense form of the copula *be*;[4] *doing* is an *-ing* participle that is monotransitive (i.e. takes a single object); *the* is a definite article; *work* is a common noun; and the specific punctuation marked that is used is a period. The manner in which the ICE tagset has been developed follows Leech's (1997: 25–6) suggestion 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").

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 British National Corpus (Garside, Leech, and Sampson 1987; Garside and Smith 1997). The TOSCA team at the University of Nijmegen has developed a tagger that can insert two tagsets: the TOSCA tagset (used to tag the Nijmegen Corpus) and the ICE tagset (Aarts, van Halteren, and Oostdijk 1996). The AUTASYS Tagger can also be used to insert the ICE tagset as well as the LOB tagset (Fang 1996). The Brill Tagger is a multi-purpose tagger that can be trained

[3] The example provided here is in a vertical format generated by the AMALGAM tagging project. This project accepts examples to be tagged by e-mail, and will tag the examples with most of the major tagsets, such as those created to tag the LOB, Brown, and ICE corpora. For more information on this project and on how the Brill Tagger was trained to insert the various tagsets, see Atwell et al. (2000) as well as the project website: http://agora.leeds.ac.uk/amalgam/amalgam/amalghome.htm. The UCREL research centre at the University of Lancaster also provides online tagging with CLAWS4 at: http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/trial.html.

[4] In the current ICE tagset, *'m* would be tagged as aux (prog, pres, encl), that is, as a present-tense progressive enclictic auxiliary.

to insert whatever tagset the user is working with into texts in English or any other language. These taggers do not exhaust the number of taggers that have been created, but provide an overview of the common taggers that have been developed to annotate the types of corpora that corpus linguists typically work with.

Taggers are of two types: rule-based or probablistic. 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 that the majority of words in English will fall into.

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 thus are candidates for disambiguation. Initially, this is done automatically by a series of "context frame rules" that look at 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 plural noun tag to this word. Although the context frame rules can disambiguate a number of tags, the process is complicated, as Francis (1979: 202) observes, by the fact that 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.

Although rule-based taggers have been largely superseded by probability-based taggers, there is one current rule-based tagger, EngCG-2 (cf. Samuelsson and Voutilainen 1997), that has been designed to overcome some of the problems inherent 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.

Because rule-based taggers rely on rules written into the tagger, most taggers developed since TAGGIT have been probabilistic in nature: 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 probabilistic taggers is that they can be trained on corpora and over time develop very high and accurate probabilities, making them quite accurate. However, even though probabilitistic and some rule-based taggers such EngCG-2 can achieve accuracy rates exceeding 95 percent, the remaining inaccuracies can be more extensive than one might think.

Kennedy (1996: 253) notes that such rates of accuracy involve "averaging": computing the success rate of tagging by combining figures for constructions (such as the definite article *the*) that have high frequencies and that can be tagged accurately almost all the time with other constructions (such as the word *once*, which can be an adverbial or a subordinating conjunction) that can be tagged accurately only 80–85 percent of the time. If a corpus is not post-edited after tagging is done (as was the case with the initial release of the British National Corpus), the rate of inaccurate tagging for certain words can be quite high. For instance, after the Wellington Corpus of Written New Zealand English was tagged with the CLAWS tagger, Kennedy (1996: 255) found that more than 20 percent of the instances of *once* were wrongly tagged. In the sentence *Once implemented, it will be with us*, *once* was tagged erroneously as an adverbial rather than a subordinating conjunction. In other instances, adverbial uses of *once* were tagged as conjunctions. Multiplied over a large corpus, error rates as high as 20 percent can lead to considerable inaccuracy, a point that users of corpora need to be aware of, especially if they intend to do an automatic analysis of the corpus without actually looking carefully at the data being analyzed. Because of the time and effort it takes to post-edit a tagged corpus, in the future we are more likely to see corpora that are released for use without any post-editing.

Because taggers cannot "sanitize" the data they have to work with, it is easy to understand why they cannot tag with complete accuracy the words of any human language. Human language is full of unusual and "idiosyncratic phenomena" (Smith 1997: 147) that taggers have to account for. For example, consider how the AMALGAM tagger analyzed the sentence *What's he want to prove?*

```
what/PRON(nom)
's/V(cop,pres,encl)
he/PRON(pers,sing)
want/V(montr,pres)
to/PRTCL(to)
prove/V(montr,infin)
?/PUNC(qm)
```

The full uncontracted form of this sentence would be *What does he want to prove*? However, because *'s* is usually a contracted form of *is*, the tagger has wrongly tagged *'s* as the present-tense enclitic form of the copula *be*. To deal with problematic cases of this type, those designing the CLAWS tagger have

developed a series of corpus "patches" (Smith 1997: 145–7): context-sensitive rules that take the output of data tagged with CLAWS and attempt to correct persistently problematic cases that have been identified by previous tagging. However, even with these rules, there remains a 2 percent rate of error in the tagging.

The tagsets used to annotate corpora are as varied as the taggers used to insert them. For instance, the Brown Corpus tagset contains seventy-seven tags (Francis and Kučera 1982), the ICE tagset 262 tags (Greenbaum and Ni 1996: 93).[5] A listing of tags for the auxiliary verb *do* illustrates the manner in which these two tagsets differ. The Brown tagset contains only three tags for *do* that specify the main surface structure forms that this verb takes:

Tag	*Form*
DO	do
DOD	did
DOZ	does

The ICE tagset, on the other hand, contains eight tags for *do* that provide a more exhaustive listing of the various forms that this auxiliary has (Greenbaum and Ni 1996: 101):

Tag	*Form*
Aux (do, infin)	Do sit down
Aux (do, infin, neg)	Don't be silly
Aux (do, past)	Did you know that?
Aux (do, past, neg)	You didn't lock the door
Aux (do, present)	I do like you
Aux (do, pres, encl)	What's he want to prove?
Aux (do, pres, neg)	You just don't understand
Aux (do, pres, procl)	D'you like ice-cream?

The differences in the tagsets reflect differing conceptions of English grammar. The Brown tagset with its three forms of *do* is based on a more traditional view of the forms that this auxiliary takes, a view that is quite viable because this tagset was designed to apply to a corpus of edited written English. The ICE tagset, in contrast, 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 low frequencies in English, but it is one of the few reference grammars of English to be based on speech as well as writing. Consequently, the ICE tagset is not just detailed but intended to account for forms of *do* that would be found in speech, such as the proenclictic form *D'you*. The more recent *Longman Grammar of Spoken and Written English* is also based on a corpus, the Longman Spoken and Written English Corpus, that was tagged by a tagger whose tagset was sensitive to distinctions made in speech and writing (cf. Biber

[5] The number of tags for the ICE tagset is approximate. As Greenbaum and Ni (1996: 93) emphasize, new constructions are always being encountered, causing additional tags to be added to the tagset.

et al. 1999: 35–8). Although it might seem that a larger tagset would lead to more frequent tagging errors, research has shown that just the opposite is true: the larger the tagset, the greater the accuracy of tagging (Smith 1997: 140–1; León and Serrano 1997: 154–7).

4.3 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 much lower accuracy rates (70–80 percent at best, depending upon how "correctness" of parsing is defined [Leech and Eyes 1997: 35 and 51, note 3]), and they require varying levels of human intervention.

Because tagging and parsing are such closely integrated processes, many parsers have taggers built into them. For instance, the Functional Dependency Grammar of English (the EngFDG parser) contains components that assign not only syntactic functions to constituents but part-of-speech tags to individual words (Voutilainen and Silvonen 1996). The TOSCA Parser has similar capabilities (Aarts, van Halteren, and Oostdijk 1996). Like taggers, parsers can be either probabilistic or rule-based, and the grammars that underlie them reflect particular conceptions of grammar, even specific grammatical theories, resulting in a variety of different "parsing schemes," that is, different systems of grammatical annotation that vary both in detail and in the types of grammatical constructions that are marked (cf. the AMALGAM "MultiTreebank" for examples of differing parsing schemes: http://www.scs.leeds.ac.uk/amalgam/amalgam/multi-tagged.html).

Although there is an ongoing debate among linguists in the field of natural language processing concerning the desirability of probabilistic vs. rule-based parsers, both kinds of parsers have been widely used to parse corpora. Proponents of probabilistic parsers have seen them as 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). This capability is largely the result of the creation of "treebanks," which aid in the training of parsers. Treebanks, such as the Lancaster Parsed Corpus and the Penn Treebank, are corpora containing 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.[6] Although many treebanks have been released and are available for linguistic analysis, their primary purpose is to train parsers to increase their accuracy.

To create grammatically analyzed corpora intended more specifically for linguistic analysis, the TOSCA Group at Nijmegen University developed the TOSCA Parser, a rule-based parser that was used to parse the Nijmegen Corpus and sections of the British component of ICE (ICE-GB). As the parse tree taken from ICE-GB in figure 4.1 illustrates, the grammar underlying the TOSCA Parser (described in detail in Oostdijk 1991) recognizes three levels of description: functions, categories, and features.

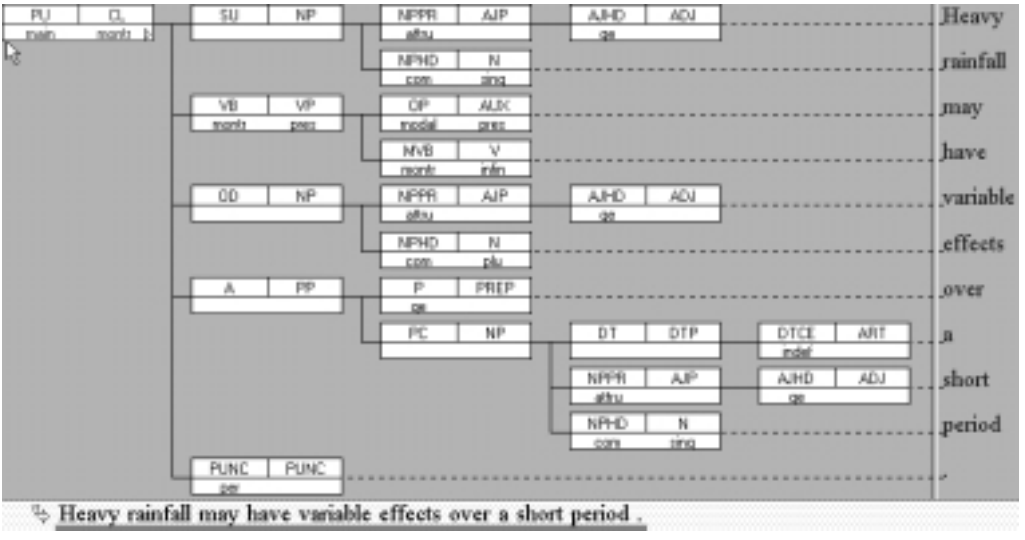


Figure 4.1 Parse tree from ICE-GB

[6] A full listing of the labels in the above example can be found in Garside, Leech, and Váradi (1992).

The first level of description – functions – is specified both within the clause and the phrase. For instance, *Heavy rainfall* is functioning as "subject" (SU) in the main clause in which it occurs; in the noun phrase itself, *Heavy* is functioning as an "adjective premodifier" (AJP). Categories are represented at both the phrase and word level: *Heavy rainfall* 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).[7]

The TOSCA Parser provides such detailed grammatical information because it was designed primarily not to create a treebank to be used to increase the accuracy of the parser but to produce a grammatically annotated corpus that "yields databases containing detailed information that may be used by linguists . . . and [that] allows for the testing of linguistic hypotheses that have been formulated in terms of the formal grammar" (Oostdijk 1991: 64). And, indeed, the two corpora parsed by the TOSCA Parser, the Nijmegen Corpus and ICE-GB, can be used with software programs that extract information from the corpora: the Linguistic Database (LDB) Program for the Nijmegen Corpus (van Halteren and van den Heuvel 1990) and the ICE Corpus Utility Program (ICECUP) for ICE-GB (cf. section 5.1.3).

While the development of taggers has "flourished" in recent years (Fang 1996: 110), parsers are still in a state of development. Taggers such as the Brill Tagger or CLAWS4 are readily available, and can be easily run on a corpus, once the corpus is in a form the tagger can accept. The output of the tagger will have to be post-edited, a process that can take some time. Nevertheless, tagging a corpus is currently a fairly common practice in corpus linguistics. Parsing, in contrast, is a much more involved undertaking, largely because of the complexity of structures that a parser is required to analyze, especially if spoken as well as written data is being parsed, and the subsequent increase in the amount of post-editing of the parsed output that needs to be done.

A parser has a much greater range of structures to analyze than a tagger: not just individual words but phrases and clauses as well. And because phrases and clauses are considerably complex in structure, there are numerous constructions, such as those that are coordinated, that are notoriously difficult to parse correctly. Figure 4.2 contains the parsed output of the sentence *The child broke his arm and his wrist and his mother called a doctor* after it was submitted to the EngFDG Parser.[8]

[7] A listing of function, category, and feature labels used in the parsing of ICE-GB can be found in the "Quick Guide to the ICE-GB Grammar" (http://www.ucl.ac.uk/english-usage/ice-gb/grammar.htm).

[8] Cf. Tapanainen and Järvinen (1997) for a more detailed discussion of the EngFDG Parser and the grammatical annotation it inserts into parsed texts such as those in figure 4.2. This is a rule-based parser influenced by the view of grammar articulated in "dependency" grammars, that is, grammars such as those described in Melčuk (1987) that instead of grouping constituents hierarchically (as parsers based on phrase structure grammars do), break each word down into its constituent parts, noting dependency relationships between constituents. The parsed output in

0
1 The the det:>2 @DN> DET SG/PL

2 child child subj:>3 @SUBJ N NOM SG

3 broke break main:>0 @+FMAINV V PAST

4 his he attr:>5 @A> PRON PERS MASC GEN SG3

5 arm arm obj:>3 @OBJ N NOM SG

6 and and @CC CC

7 his he attr:>8 @A> PRON PERS MASC GEN SG3

8 wrist wrist @**SUBJ** N NOM SG @**OBJ** N NOM SG @PCOMPL-S N NOM SG @A>N NOM SG

9 and and cc:>3 @CC CC

10 his he attr:>11 @A> PRON PERS MASC GEN SG3

11 mother mother subj:>12 @SUBJ N NOM SG

12 called call cc:>3 @+FMAINV V PAST

13 a a det:>14 @DN> DET SG

14 doctor doctor obj:>12 @OBJ N NOM SG

Figure 4.2 Parsed output from the EngFDG Parser

The example in figure 4.2 contains the coordinator *and* conjoining two noun phrases, *his arm and his wrist*, as well as two main clauses. Because coordinators such as *and* can be used to conjoin phrases as well as clauses, it can be difficult for parsers to distinguish phrasal from clausal coordination. In the example in figure 4.2, the parser is unable to determine whether *his wrist* is coordinated with *his arm* or *his mother*. Therefore, *his wrist* is assigned two function labels: subject (@SUBJ) to indicate that *his wrist* and *his mother* are coordinated noun phrases functioning as subject of the second clause; and object (@OBJ), which is actually the correct parse, to indicate that *his arm* and *his wrist* are coordinated noun phrases functioning as object of the verb *broke* in the first clause. The difficulty that parsers have with coordinated constructions is further compounded by the fact that coordination is so common (the one-million-word ICE-GB contains over 20,000 instances of *and*), increasing the likelihood of multiple parses for coordinated constructions.

An additional level of complexity is introduced if a parser is used to parse spoken as well as written data. In this case, the parser has to deal with the sheer ungrammaticality of speech, which poses major challenges for parsers, especially those that are rule-based. For instance, the utterance below, taken from

figure 4.2 was obtained by using the demonstration version of the EngFDG parser available on the Conexor website at: http://www.conexor.fi/testing.html#1.

ICE-USA, is typical of the "ungrammatical" nature of speech: it has numerous false starts, leading to a series of partially formed sentences and clauses:

> but anyway you were saying Peggy that well I was asking you the general question what um how when you've taken courses in the linguistics program have you done mainly just textbook stuff or have you actually this is actual hardcore linguistic scholarship
>
> (cited in Greenbaum 1992: 174)

In a rule-based parser, it is impossible to write rules for utterances containing repetitions, since such utterances have no consistent, predictable structure and thus conform to no linguistic rule. In addition, while a written text will contain clearly delineated sentences marked by punctuation, speech does not. Therefore, with examples such as the above, it is difficult to determine exactly what should be submitted to the parser as a "parse unit": the stretch of language the parser is supposed to analyze (Gerald Nelson, personal communication).

The increased complexity of structures that a parser has to analyze leads to multiple parses in many cases, making the process of tag disambiguation more complex than it is with the output of a tagger. With tagging, disambiguation involves selecting one tag from a series of tags assigned to an individual word. With parsing, in contrast, disambiguation involves not single words but higher-level constituents; that is, instead of being faced with a situation where it has to be decided whether a word is a noun or a verb or adjective, the analyst has to select the correct parse tree from any number of different parse trees generated by the parser. In using the TOSCA parser, for instance, Willis (1996: 238) reports that after running the parser on a sentence, if a single correct parse tree was not generated, he had to select from a dozen to (on some occasions) 120 different parse trees. There is software to help in this kind of post-editing, such as ICE Tree (described at: http://www.ucl.ac.uk/english-usage/ice-gb/icetree/download.htm), a general editing program that was designed to aid in the post-editing of corpora tagged and parsed for the ICE project, and the Nijmegen TOSCA Tree editor (described at: http://lands.let.kun.nl/TSpublic/tosca/page4.html), which can also be used to post-edit the output of the TOSCA parser. But still, any time there is manual intervention in the process of creating a corpus the amount of time required to complete the corpus multiplies exponentially.

To increase the accuracy of parsing and therefore reduce the amount of post-editing that is required, parsers take differing approaches. The TOSCA Parser requires a certain amount of manual pre-processing of a text before it is submitted to the parser: syntactic markers are inserted around certain problematic constructions (cf. Oostdijk 1991: 263–7; Quinn and Porter 1996: 71–4), and spoken texts are normalized. To ensure that coordinated structures are correctly parsed, their beginnings and ends are marked. In the sentence *The child broke his arm and his wrist and his mother called a doctor* (the example cited above), markers would be placed around *his arm and his wrist* to indicate that these two noun phrases are coordinated, and around the two main clauses that are

coordinated to indicate that the coordinator *and* following *wrist* conjoins two clauses. Other problematic constructions that had to be marked prior to parsing were noun phrase postmodifiers, adverbial noun phrases, appositive noun phrases, and vocatives.

To successfully parse instances of dysfluency in speech, the parser requires that, in essence, an ungrammatical utterance is made grammatical: markup has to be manually inserted around certain parts of an utterance that both tells the parser to ignore this part of the utterance and creates a grammatically well-formed structure to parse. In the example below, the speaker begins the utterance with two repetitions of the expression *can I* before finally completing the construction with *can we.*

can I can I can we take that again (ICE-GB S1A-001–016)

To normalize this construction, markup is inserted around the two instances of *can I* ("<->" and "</->") that tells the parser to ignore these two expressions and parse only *can we take that again*, which is a grammatically well formed:

<sent> <}> <-> Can I can I </-> <=> can we </=> </}> take that again <$?>

Arguably, normalization compromises the integrity of a text, especially if the individual doing the normalization has to make decisions about how the text is to be edited. And because normalization must be done manually, it can be time-consuming as well. But normalization has advantages too. In constructions such as the above, if the repetitions are not annotated, they will be included in any lexical analysis of the corpus, leading to inaccurate word counts, since *can*, for instance, will be counted three times, even though it is really being used only once. In addition, even though repetitions and other features of speech are excluded from analysis, if they are properly annotated, they can be recovered if, for instance, the analyst wishes to study false starts.

Even though parsing a corpus is a formidable task, there do exist a number of parsed corpora available for linguistic research, though many of them are relatively short by modern standards. In addition to the parsed corpora described above – the Nijmegen Corpus (130,000 words), the Lancaster Parsed Corpus (140,000 words), and ICE-GB (one million words) – there are also the Lancaster/IBM Spoken English Corpus (52,000 words), the Polytechnic of Wales Corpus (65,000 words), the Penn–Helsinki Parsed Corpus of Middle English (1.3 million words), and the Susanne Corpus (128,000 words). The Penn Treebank is larger (3,300,000 words) than all of these corpora, but aside from the texts taken from the Brown Corpus, this corpus is not balanced, consisting primarily of reportage from Dow Jones news reports. The ENGCG Parser (an earlier version of the EngFDG Parser) has been used to parse sections of the Bank of English Corpus (over 100 million words as of 1994), with the goal of parsing the entire corpus (cf. Järvinen 1994 for details).

4.4 Other types of tagging and parsing

The previous sections described the tagging of words and the parsing of constituents into syntactic units. While these processes are very well established in the field of corpus linguistics, there are other less common types of tagging and parsing as well.

4.4.1 Semantic tagging

Semantic tagging involves annotating a corpus with markup that specifies various features of meaning in the corpus. Wilson and Thomas (1997) describe a number of systems of semantic tagging that have been employed. In each of these systems, words in corpora are annotated with various schemes denoting their meanings. For instance, in one scheme, each word is assigned a "semantic field tag" (Wilson and Thomas 1997: 61): a word such as *cheeks* is given the tag "Body and Body Parts," a word such as *lovely* the tag "Aesthetic Sentiments," and so forth. Schemes of this type can be useful for the creation of dictionaries, and for developing systems that do "content analysis," that is, search documents for particular topics so that users interested in the topics can have automatic analysis of a large database of documents.

4.4.2 Discourse tagging

In addition to annotating the meaning of words in corpora, semantic systems of tagging have also looked at such semantic phenomena as anaphora, the chaining together of co-referential links in a text. This is a type of discourse tagging, whereby features of a text are annotated so that analysts can recover the discourse structure of the text. Rocha (1997) describes a system he developed to study "topics" in conversations. In this system, Rocha (1997) developed an annotation scheme that allowed him to classify each anaphor he encountered in a text and mark various characteristics of it. For instance, each anaphor receives annotation specifying the type of anaphor it is: personal pronouns are marked as being either subject pronouns (SP) or object pronouns (OP); demonstratives as (De); possessives as (PoP); and so forth (Rocha 1997: 269). Other information is also supplied, such as whether the antecendent is explicit (ex_) or implicit (im_). Rocha (1997) used this scheme to compare anaphor resolution in English and Portuguese, that is, what chain of "links" are necessary to ultimately resolve the reference of any anaphor in a text. Rocha's (1997: 277) preliminary results show that most anaphors are personal pronouns, and their reference is predominantly explicit.

4.4.3 Problem-oriented tagging

De Haan (1984) coined the term "problem-oriented" tagging to describe a method of tagging that requires the analyst to define the tags to be used

and to assign them manually to the constructions to be analyzed. For instance, Meyer (1992) used this method to study appositions in the Brown, London–Lund, and Survey of English Usage Corpora. Each apposition that Meyer (1992) identified was assigned a series of tags. An apposition such as *one of my closest friends, John* in the sentence *I called one of my closest friends, John* would be assigned various tag values providing such information as the syntactic form of the apposition, its syntactic function, whether the two units of the apposition were juxtaposed or not, and so forth. Within each of these general categories were a range of choices that were assigned numerical values. For instance, Meyer (1992: 136–7) found that the appositions in the corpora he analyzed had seventy-eight different syntactic forms. The above apposition had the form of an indefinite noun phrase followed by a proper noun, a form that was assigned the numerical value of (6). By assigning numerical values to tags, the tags could be subjected to statistical analysis.

Because problem-oriented tagging has to be done manually, it can be very time-consuming. However, there is a software program, PC Tagger, that can be used to expedite this type of tagging (Meyer and Tenney 1993). The advantage of problem-oriented tagging is that the analyst can define the tags to be used and is not be constrained by someone else's tagset. In addition, the tagging can be quite detailed and permit grammatical studies that could not be carried out on a corpus that has only been lexically tagged.

The term problem-oriented tagging suggests that this type of analysis involves mainly tagging. But it is really a process of parsing as well: constituents larger than the word are assigned syntactic labels.

4.5 Conclusions

In recent years, corpus linguists of all persuasions have been actively involved in developing systems of annotation for corpora. The Linguistic Data Consortium (LDC) at the University of Pennsylvania, for instance, has a whole web page of links to various projects involved in either annotating corpora or developing tools to annotate them ("Linguistic Annotation": http://www.ldc.upenn.edu/annotation/). This is a significant and important development in corpus linguistics, since an annotated corpus provides the corpus user with a wealth of important information.

But while it is important that systems of annotation are developed, it is equally important that corpus linguists develop tools that help reduce the amount of time it takes to annotate corpora and that help users understand the complex systems underlying many annotation systems. The Text Encoding Initiative (TEI) has developed a comprehensive system of structural markup. However, it is very time-consuming to insert the markup, and many users may have difficulty understanding how the TEI system actually works. Likewise, parsing a corpus is a very involved process that requires the corpus creator not just to spend

considerable time post-editing the output of the parser but to have a conceptual understanding of the grammar underlying the parser. Because the process of annotating corpora involves the manual intervention of the corpus creator at various stages of annotation, in the foreseeable future, annotating a corpus will continue to be one of the more labor-intensive parts of creating a corpus.

Study questions

1. Why is it necessary for a corpus in ASCII text format to contain structural markup? What kinds of features would users of such a corpus not be able to recover if the samples within the corpus did not contain structural markup?
2. How would a probabilistic tagger determine that in the sentence *The child likes to play with his friends*, the word *play* is a verb rather than a noun?
3. Make up a short sentence and submit it to the online taggers below:
   CLAWS part-of-speech tagger:
   http://www.comp.lancs.ac.uk/computing/research/ucrel/claws/trial.html
   EngCG-2:
   http://www.conexor.fi/testing.html#1
   Do the two taggers assign similar or different part-of-speech labels to your sample sentence? Briefly summarize the similarities and differences.
4. Why is parsing a much more difficult and involved process than tagging?

5 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. It begins with a discussion of how to frame a research question so that from the start, the analyst has a clear "hypothesis" to test out in a corpus and avoids the common complaint that many voice about corpus-based analyses: that many such analyses do little more than simply "count" linguistic features in a corpus, paying little attention to the significance of the counts. The next sections describe the process of doing a corpus analysis: how to determine whether a given corpus is appropriate for a particular linguistic analysis, how to extract grammatical information relevant to the analysis, how to create data files for recording the grammatical information taken from the corpus, and how to determine the appropriate statistical tests for analyzing the information in the data files that have been created.

To illustrate the process of analyzing a corpus, this chapter will focus on a specific corpus analysis: the study of a particular grammatical construction, termed the "pseudo-title" (Bell 1988), in various components of the International Corpus of English (ICE). Pseudo-titles are constructions such as *linguist* in *linguist Noam Chomsky* that occur in positions in which we normally find titles like *Professor* or *Doctor*. However, unlike titles, pseudo-titles do not function as honorifics (markers of deference and respect) but instead provide descriptive information about the proper nouns that they precede (e.g. that Noam Chomsky is a "linguist"). The "case study" approach is taken in this chapter to avoid using disparate and unrelated examples to illustrate the process of doing a corpus analysis and to demonstrate that each stage of corpus analysis is related.[1]

[1] For additional case studies, see Biber, Conrad, and Reppen (1998).

5.1 The pseudo-title in English: framing a research question

To begin to frame a research question about pseudo-titles, it is first of all necessary to understand the history of the construction and how it has evolved in English.

Pseudo-titles are thought to have originated in American English, specifically in journalistic writing. Quirk et al. (1985: 276, note) refer to pseudo-titles as characteristic of "Timestyle," a style of journalistic writing popularized by *Time* magazine. Although pseudo-titles may have originated in American press writing, they can currently be found in the press writing of both British and New Zealand English. However, as Rydén (1975), Bell (1988), and Meyer (1992) have documented, while pseudo-titles are not stylistically marked in American English – they occur widely in all types of American newspapers and magazines – they are stigmatized in British English: their use is avoided in more formal newspapers such as *The Times* and the *Guardian*, and they are found mainly in newspapers associated with what is now termed "tabloid" journalism.

The stigma against pseudo-titles in British English is well documented in style guides for British newspapers. For instance, in *The Independent Style Manual*, it is remarked that pseudo-titles should be avoided "because people's jobs are not titles" (p. 6). In newspapers prohibiting the use of pseudo-titles, the information contained in the pseudo-title is instead expressed in an equivalent appositional construction (e.g. *the linguist Noam Chomsky* or *Noam Chomsky, a linguist*). But while style guides may prohibit certain usages such as the pseudo-title, it is an open question whether practice follows prescription, and a well-designed corpus study can be used to examine the relationship between prescription and practice. Bell (1988), for instance, studied 3,500 instances of pseudo-titles taken from a large corpus of British, American, and New Zealand newspapers. Using this corpus, Bell (1988) was able to confirm that while pseudo-titles were quite common in American press writing, they were absent from British newspapers that prohibited their usage. Additionally, Bell (1998) was able to document that New Zealand press writing had moved away from the British norm for pseudo-titles to the American norm, that is, that over the years, pseudo-titles had gained in prominence and acceptability in New Zealand press writing. This change in usage norms, Bell (1988: 326–7) argues, reflects the fact that the colonial influence of British culture on New Zealand culture had "been weakening" in recent years, and that New Zealand English was "refocusing towards the United States . . . ," a change that has led New Zealand English closer to the American norm for pseudo-titles. The size of Bell's (1988) corpus also enabled him to study the linguistic structure of pseudo-titles, and to document, for instance, the extent that newspapers favored pseudo-titles over equivalent appositional structures, and to show that there were certain linguistic structures that favored or disfavored the use of a pseudo-title. For instance, he

found that pseudo-titles favored minimal post-modification. Consequently, a construction such as *lawyer Frederick Smith* was more common than ?*lawyer for Hale, Brown, and Jones Frederick Smith*.

To evaluate whether Bell's (1988) study of pseudo-titles uses a corpus to answer valid research questions, and to determine whether his study warrants further corpus studies of pseudo-titles, it is important to understand the kinds of linguistic evidence that corpora are best able to provide. Since the Chomskyan revolution of the 1950s, linguistics has always regarded itself as a science in which "empirical" evidence is used to advance linguistic theory. For the Chomskyan linguist, this evidence was often obtained through "introspection": the gathering of data based on the linguist's own intuitions. Because many corpus linguists have found this kind of evidence limiting (cf. section 1.1), they have turned to the linguistic corpus as a better source of empirical evidence – for "real" rather than "contrived" examples of linguistic constructions and for statistical information on how frequently a given linguistic construction occurs in a corpus.

In turning to the corpus for evidence, however, many corpus linguists have regarded the gathering of evidence as a primary goal: in some corpus studies, it is not uncommon to see page after page of tables with statistical information on the frequency of grammatical constructions, with little attention paid to the significance of the frequencies. Corpus-based research of this nature, Aarts (2001: 7) notes,

> invariably elicits a "so what" response: so what if we know that there are 435 instances of the conjunction "because" in a particular category of written language, whereas there are only 21 instances in conversations? So what if we are told that subordination is much more common in women's speech than in men's speech?

To move beyond simply counting features in a corpus, it is imperative before undertaking a corpus analysis to have a particular research question in mind, and to regard the analysis of a corpus as both "qualitative" and "quantitative" research – research that uses statistical counts or linguistic examples to test a clearly defined linguistic hypothesis.

In using a corpus to study pseudo-titles, Bell (1988) has obviously done more than simply count the number of pseudo-titles in his corpus. He has used a corpus to, for instance, describe the particular linguistic structures that pseudo-titles favor, and he has used frequency counts to document the increased prevalence of pseudo-titles in New Zealand press writing. Because Bell's (1988) study suggests that pseudo-titles may be spreading beyond British and American English to other regional varieties of English, his findings raise additional questions worthy of further investigation of corpora:

(a) To what extent have pseudo-titles spread to other varieties of English, and in these varieties, is the American norm followed or the British norm?
(b) Do pseudo-titles have the same structure in these varieties that they have in British, American, and New Zealand English?

(c) To what extent do newspapers in these varieties prefer pseudo-titles over equivalent appositional structures?

The first step in addressing these questions is to determine whether suitable corpora exist for answering them.

5.2 Determining whether a corpus is suitable for answering a particular research question

Bell's (1988) study of pseudo-titles was based on a large corpus of British, American, and New Zealand press writing that Bell assembled himself. Although this is the ideal way to ensure that the corpus being used is appropriate for the analysis being conducted, this approach greatly increases the work that the corpus analyst must do, since creating one's own corpus prior to analyzing it is obviously a large undertaking. It is therefore most desirable to work with a corpus already available not just to decrease the work time but to add to the growing body of work that has been based on a given corpus. The Brown Corpus, for instance, has been around for close to forty years and has become a type of "benchmark" for corpus work: a corpus on which a large amount of research has been conducted over the years – research that has generated results that can be compared and replicated on other corpora, and that has added greatly to our knowledge of the types of written English included in the Brown Corpus.

To study the spread of pseudo-titles to various regional varieties of English, it is necessary not only to find corpora of differing varieties of English but to make sure that the corpora selected are as "comparable" as possible, that is, that they contain texts that have been collected during a similar time period, that represent similar genres, that are taken from a variety of different sources, and that are of the appropriate length for studying pseudo-title usage. In other words, the goal is to select corpora that are as similar as possible so that extraneous variables (such as texts covering different time periods) are not introduced into the study that may skew the results.

Since pseudo-titles have been extensively studied in British, American, and New Zealand English, it might appear that corpora of these varieties need not be sought, since analyzing these varieties again will simply repeat what others have already done. However, there are important reasons why corpora of British, American, and New Zealand English ought to be included in the study. First, it is always desirable to have independent confirmation of results obtained in previous studies. Such confirmation ensures that the results have validity beyond the initial study from which they were obtained. Second, the previous studies of pseudo-titles were conducted on very different kinds of corpora. Bell's (1988) study was conducted on a privately created corpus; Meyer's (1992) study was based on the Brown Corpus and the London Corpus, corpora consisting of samples of written British and American English collected as far back as the late

1950s. Comparing very different corpora increases the potential that extraneous variables will influence the results. For instance, because Bell's (1988) study suggests that pseudo-title usage seems to be increasing over time, it is important to study corpora collected during a similar time period so that the variable of time does not distort the results.

There are several well-established corpora that potentially qualify as resources for studying pseudo-title usage: the Brown and LOB corpora of American and British English, respectively (and their updated counterparts: the Freiburg Brown Corpus [FROWN] and Freiburg LOB Corpus [FLOB]); the Wellington Corpus of New Zealand English; and the Kolhapur Corpus of Indian English. Because each of these corpora has a similar design (e.g. comparable written genres divided into 2,000-word samples), they are very suitable for comparison. However, the Wellington Corpus contains texts collected at a later date than the other corpora: 1986-90 as opposed to 1961. FLOB and FROWN contain texts from 1991. These corpora thus introduce a time variable that is less than desirable. If these corpora are eliminated from consideration, only three corpora remain (Brown, LOB, and Kolhapur), and only one of these corpora – the Kolhapur Corpus of Indian English – represents a variety of English in which pseudo-titles have never before been studied. It is therefore obvious that a different set of corpora are necessary if any information is to be obtained on the spread of pseudo-titles to other varieties of English than American and British English.

A better choice of corpora to consider are those included in the International Corpus of English (ICE). The regional components of the ICE contain not just identical genres but samples of speech and writing collected during a similar time-frame (1990–present). In addition, ICE corpora offer numerous regional varieties – not just British, American, and New Zealand English but Philippine, Singaporean, Jamaican, and East African English.[2] The ICE corpora therefore offer a potentially useful dataset for studying pseudo-titles. But to determine the full potential of the ICE corpora, two further points must be considered: whether the components of ICE selected for analysis provide a representative sampling of the kind of writing that pseudo-titles occur in, and if the components do, whether the samples selected from the corpora are of the appropriate length to find a sufficient number of pseudo-titles and corresponding appositives.

In his study of pseudo-titles, Bell (1988: 327) restricted his analysis to pseudo-titles occurring in "hard news reporting" and excluded from consideration other types of press writing, such as editorials or feature writing. Because each ICE corpus contains twenty samples of press reportage as well as twenty samples of broadcast news reports, it will be quite possible to base an analysis of pseudo-titles on the kinds of press reportage that Bell (1988) analyzed in his study. The next question, however, is whether the analysis should be based on all

[2] The ICE Project contains other varieties of English than these, but these were the only varieties available at the time this study was conducted.

forty texts from each ICE corpus, or some subset of these texts. There are a number of considerations that will affect choices of this nature. First of all, if the construction to be analyzed has to be extracted by manual analysis, then it becomes imperative to use as short a corpus as possible, since extensive manual analysis will add to the length of time a given study will take. As will be demonstrated in section 5.3, while pseudo-titles can be automatically extracted from the British component of ICE because it is fully tagged and parsed, they must be retrieved manually from the other components, since unlike, say, regular verbs in English, which have verb suffixes such as *-ing* or *-ed*, pseudo-titles have no unique marker that can be searched for. Therefore, to minimize the analysis time, it is advisable to examine only a subset of the eligible texts in the various ICE components.

One possibility is to analyze ten samples of press reporting and ten samples of broadcast news from each ICE corpus. Another possibility is to analyze only the twenty samples of press reportage from each corpus, or only the twenty samples of broadcast news. Ultimately, it was decided to restrict the sample to the twenty samples of press reportage, primarily because broadcast news is a mixed category, containing not just scripted monologues but interviews as well. Since interviews are not strictly "hard news reporting," they are less likely to contain pseudo-titles.[3] Moreover, since the goal of the study is to determine how widespread pseudo-titles are in press reportage, it is quite sufficient to attempt to investigate as many different newspapers as possible from the various ICE corpora, to determine how common pseudo-title usage is.

Having restricted the number of texts to be examined, it is next necessary to examine the individual samples from each component of ICE to ensure that the samples represent many different newspapers rather than a few newspapers. In other words, if a given component of ICE contains twenty samples taken from only five different newspapers, the samples to be investigated are less than representative, since they provide information only on how five different newspapers use pseudo-titles. With a sample this small, it will be difficult to make strong generalizations about the usage of pseudo-titles in general within the press writing of a given variety of English. Table 5.1 lists the number of different newspapers represented in the samples selected for study from the various components of ICE.

As table 5.1 illustrates, the components vary considerably in terms of the number of different newspapers that are represented. While ICE-USA contains

[3] Robert Sigley (personal communication) notes that pseudo-titles would be perfectly appropriate in a broadcast interview as a way of introducing an interviewee: "We will be speaking with Boston Globe Editor Robert Storrin." In ICE-GB, the only pseudo-title found in a broadcast interview (*designer* in *designer Anthony Baker*) occurred in an interviewee's description of a play he recently saw:

> The staging by Francisco Negrin and *designer Anthony Baker* wittily plays off the opera house and its conventions finally allowing us to glimpse a three tier gilt and scarlet plush auditorium into which the cast make their escape (ICE-GB:S1B-044 #2:1:A; emphasis added)

Table 5.1 *Number of different newspapers included within the various ICE components*

Component	Total
Great Britain	15
United States	20
New Zealand	12
Philippines	10
Jamaica	3
East Africa	3
Singapore	2

twenty different newspapers, ICE-East Africa contains only three. The relatively few newspapers represented in some of the components is in part a reflection of the fact that in an area such as East Africa, in which English is a second language, there exist only a few English language newspapers. Therefore, three newspapers from this region is a representative sample. Still, the results obtained from such a small sample will have to be viewed with some caution.

Because all ICE corpora contain 2,000-word samples, it is also important to consider whether text fragments, rather than entire texts, will yield enough examples of pseudo-titles and, additionally, whether pseudo-titles serve some kind of discourse function, necessitating that a whole newspaper article be analyzed as opposed to simply a short segment of it.[4] One way to answer these questions is to do a pilot study of a few randomly selected samples taken from each ICE component. Table 5.2 lists the number of pseudo-titles and corresponding appositive structures occurring in nine samples from various randomly selected ICE corpora.

Table 5.2 demonstrates that even though the frequency of the various structures being investigated differs in the nine newspapers, the structures do occur with some regularity in the 2,000-words samples: as many as sixteen times in one sample from ICE-New Zealand to as few as five times in one sample from ICE-Philippines. Of course, ultimately statistical tests will have to be applied to determine the true significance of these numbers. But at this stage, it seems evident that pseudo-titles are such a common grammatical construction that adequate numbers of them can be found in short 2,000-word excerpts.

One drawback of analyzing text fragments is that certain discourse features cannot be studied in them because only partial texts (rather than whole texts) are available for analysis. In particular, if one is investigating a construction that contributes to the structure of a given text, then a text fragment will prove insufficient for carrying out such a study. Because pseudo-titles serve mainly

[4] Many ICE samples are actually lengthier than 2,000 words. However, the text of a sample beyond 2,000 words is marked as "extra-corpus text" and for this study was excluded from analysis.

Table 5.2 *The number of pseudo-titles and corresponding appositives in selected newspapers from various ICE corpora*

	Pseudo-title	Equivalent appositive	Total
Independent (ICE-GB)		9	9
Guardian (ICE-GB)		7	7
Press (ICE-NZ)		13	13
Daily Nation (ICE-EA)	4	9	13
NY Times (ICE-US)	2	6	8
Jamaica Herald (ICE-JA)	2	4	6
Phil. Star (ICE-PH)	4	1	5
Manila Standard (ICE-PH)	11	0	11
Dominion (ICE-NZ)	16	0	16

to provide descriptive information about individuals, they play no role in the overall structure or organization of a text. Therefore, it is perfectly justifiable to study pseudo-title usage in the 2,000-word excerpts occurring in samples from ICE corpora, particularly if the goal is to study the structure and frequency of pseudo-titles and corresponding appositives in various national varieties of English.

5.3 Extracting information from a corpus

After a research question has been framed and a corpus selected to analyze, it is next necessary to plan out exactly what kinds of grammatical information will be extracted from the corpus, to determine how this information will be coded and recorded, and to select the appropriate software that can most efficiently assist in finding the construction being investigated in the corpus being studied.

5.3.1 Defining the parameters of a corpus analysis

The first step in planning a corpus analysis is the development of a clear working definition of the grammatical construction(s) to be studied. If a corpus analysis is begun without an adequate definition of terms, the analyst runs the risk of introducing too much inconsistency into his/her analysis. In the case of pseudo-titles, this becomes an especially crucial issue because research has shown that there is not always a clear boundary between a full title and a pseudo-title. If some kind of working definition of each kind of title is not determined prior to analysis, it is highly likely that as the analysis is conducted what is counted as a pseudo-title will vary from instance to instance, and as a consequence, considerable inconsistency will result that will seriously

compromise the integrity of the findings. Fortunately, Bell (1988) has specified quite precisely the kinds of semantic categories into which full titles can be categorized, and his categories can be used to distinguish what counts as a title from what counts as a pseudo-title:

> Professional (*Doctor, Professor*)
> Political (*President, Chancellor, Senator*)
> Religious (*Bishop, Cardinal, Mother*)
> Honors (*Dame, Earl, Countess*)
> Military (*General, Corporal*)
> Police (*Commissioner, Constable, Detective-Sergeant*)
> Foreign (*Monsieur, Senorita*) (Bell 1988: 329)

In the present study, then, a pseudo-title will be defined as any construction containing an initial unit (such as *city employee* in *city employee Mark Smith*) that provides descriptive information about an individual and that cannot be classified into any of the above semantic classes.

Because pseudo-titles will be compared to corresponding appositives, it is also necessary to have a working definition of the types of appositives that will be considered "equivalent" to a pseudo-title. Bell (1988) claims that pseudo-titles and equivalent appositives are related through a process he terms "determiner deletion." Thus, for Bell (1988: 328), a pseudo-title such as *fugitive financier Robert Vesco* would be derived from an appositive containing the determiner *the* in the first unit: *the fugitive financier Robert Vesco*. However, if one's goal is to study the relationship between the use of pseudo-titles and all possible equivalent appositives, then a wider range of appositive structures needs to be investigated, largely because there exist corresponding structures where the unit of the appositive equivalent to the pseudo-title occurs after, not before, the noun phrase it is related to, and in such constructions, a determiner is optional: *Robert Vesco, [the] fugitive financier*. In the current study, then, a corresponding appositive will include any appositive that is related to a pseudo-title through the process "systematic correspondence" (Quirk et al. 1985: 57f.), a process that specifies that two constructions are equivalent if they contain roughly the same "lexical content" and have the same meaning. Although constructions such as *fugitive financier Robert Vesco* and *Robert Vesco, fugitive financier* certainly differ in emphasis and focus, they are close enough in form and meaning that they can be considered equivalent structures.

Of course, a working definition is just that: a preliminary definition of a grammatical category subject to change as new data are encountered. And as anyone who has conducted a corpus study knows, once "real" data are examined, considerable complications can develop. As the study of pseudo-titles was conducted, examples were found that required modification of the initial definition (cf. section 5.3.3 for a discussion of the actual methods that were employed to locate pseudo-titles). For instance, the construction *former Vice President Dan Quayle* (ICE-USA) contains an initial unit, *former Vice President*, that

is semantically a "Political" designation and therefore qualifies as a title; moreover, *Vice President* is capitalized, a common orthographic characteristic of titles. But the inclusion of the adjective *former* before *Vice President* suggests that the focus here is less on "honoring" Dan Quayle and more on his previous occupation as Vice President of the United States. Consequently, this construction (and others like it) was counted as a pseudo-title, even though *Vice President* has semantic characteristics normally associated with titles.

Decisions of this nature are common in any corpus analysis that involves the close examination of data, and there are basically two routes that the corpus analyst can take to decide how to classify problematic data. As was done above, if enough evidence exists to place a construction into one category rather than another, then the construction can be classified as "x" rather than "y," even though the construction may have characteristics of both "x" and "y." Alternatively, the analyst can create an ad hoc category in which constructions that cannot be neatly classified are placed. Either approach is acceptable, provided that whichever decision is made about the classification of a particular grammatical construction is consistently applied throughout a given corpus analysis, and that the decisions that are made are explicitly discussed in the article or book in which the results of the study are reported. Moreover, if an ad hoc category is created during a corpus analysis, once the analysis is completed, all examples in the category should be examined to determine whether they have any characteristics in common that might lead to additional generalizations.

5.3.2 Coding and recording grammatical information

Once key terms and concepts have been defined, it is next necessary to decide what grammatical information needs to be extracted from the corpus being examined to best answer the research questions being investigated, and to determine how this information can best be coded and recorded. To find precise answers to the various linguistic questions posed about pseudo-titles in section 5.1, the following grammatical information will need to be obtained:

Information A: To determine how widespread pseudo-title usage has become, and to reveal which newspapers permit the use of pseudo-titles and which prefer using only equivalent appositive constructions, it will be necessary to obtain frequency counts of the number of pseudo-titles and equivalent appositives occurring in the various regional varieties of ICE being investigated.

Information B: To determine how common pseudo-titles are in newspapers that allow their use, it will be important to examine the role that style plays in the use of pseudo-titles and equivalent appositives, and to ascertain whether given the choice between one construction and the other, a particular newspaper will choose a pseudo-title over a corresponding appositive construction. To study stylistic choices, three types of correspondence relationships will be coded: those in which an appositive can be directly converted into a pseudo-title:

(1) Durk Jager, executive vice president (ICE-USA) →
executive vice president Durk Jager

those in which only a determiner needs to be deleted for the appositive to be converted into a pseudo-title:

(2) the Organising Secretary, Mr Stephen Kalonzo Musyoka (ICE-East Africa) →
Organising Secretary Mr Stephen Kalonzo Musyoka

and those in which only part of the appositive can be converted into a pseudo-title:

(3) Peter Houliston, Counsellor and Programme Director of Development at the Canadian High Commission (ICE-Jamaica) →
Counsellor and Programme Director Peter Houliston

Information C: To investigate why only part of the appositive can form a pseudo-title in constructions such as (3), it is necessary to study the linguistic considerations that make conversion of the entire unit of an appositive to a pseudo-title stylistically awkward. Bell (1988: 336) found that an appositive construction was favored (4a and 5a) over an equivalent pseudo-title (4b and 5b) if the unit that could potentially become a pseudo-title contained either a genitive noun phrase (e.g. *bureau's* in 4a) or postmodification of considerable complexity (e.g. *. . . of the CIA station in Rome in the 1970s* in 5a):

(4) a. the bureau's litigation and prosecution division chief Osias Baldivino (ICE-Philippines)
b. ?*bureau's litigation and prosecution division chief Osias Baldivino
(5) a. Ted Shackley, deputy chief of the CIA station in Rome in the 1970s (ICE-GB:W2C-010 #62:1)
b. ?deputy chief of the CIA station in Rome in the 1970s Ted Shackley

Information D: It was decided to record the length of a pseudo-title after a casual inspection of the data seemed to suggest that in the Philippine component there was greater tolerance for lengthy pseudo-titles (6) than for what seemed to be the norm in the other varieties: pseudo-titles of only a few words in length (7).

(6) Salamat and Presidential Adviser on Flagship Projects in Mindanao Robert Aventajado (ICE-Philippines)
(7) Technology editor Kenneth James (ICE Singapore)

Often after a corpus analysis has begun, new information will be discovered that is of interest to the study being conducted and that really needs to be recorded. Usually such discoveries can be easily integrated into the study, provided that the analyst has kept accurate record of the constructions that have already been recorded, and has coded the data in a manner that allows for the introduction of new grammatical information into the coding scheme.

Data can be coded and recorded in numerous ways. Because the goal of the current study is to record specific information about each construction being studied, it was decided to use a type of manual coding (or tagging) called "problem-oriented tagging" (De Haan 1984; cf. also section 4.4.3). This method of tagging allows the analyst to record detailed information about each grammatical construction under investigation. Table 5.3 outlines the coding system used for recording the grammatical information described in (A)–(D) above.

The coding scheme in table 5.3 allows for every pseudo-title or equivalent appositive construction to be described with a six-sequence series of numbers that not only label the construction as being a pseudo-title or appositive but describe the regional variety in which the construction was found; the particular sample from which it was taken; the type of correspondence relationship existing between an appositive and potential pseudo-title; the form of the construction; and the length of the pseudo-title or the unit of the apposition that could potentially be a pseudo-title. Coding the data this way will ultimately allow for the results to be viewed from a variety of different perspectives. For instance, because each construction is given a number identifying the regional variety in which it occurred, it will be possible to compare pseudo-title usage in, say, British English and Singapore English. Likewise, each sample represents a different newspaper. Therefore, by recording the sample from which a construction was taken, it will be possible to know which newspapers permit pseudo-title usage, and which do not, and the extent to which newspapers use pseudo-titles and equivalent appositives similarly or differently.

According to the scheme in table 5.3, a construction such as *financial adviser David Innes* (ICE-GB:W2C-009 #41:2) would be coded in the following manner:

Country: Great Britain	6
Sample: W2C-009	9
Type: pseudo-title	1
Correspondence relationship	4
Form: simple NP	1
Length: two words	2

The advantage of using a coding scheme such as this is that by assigning a series of numerical values to each construction being studied, the results can be easily exported into any statistical program (cf. section 5.4 for a discussion of the kinds of statistical programs into which quantitative data such as this can be exported).

The disadvantage of using numerical sequences of this nature is that they increase the likelihood that errors will occur during the coding process. If, for instance, a pseudo-title is mistakenly tagged as an apposition, it is difficult for the analyst to recognize this mistake when each of these categories is given the arbitrary number "1" and "2," respectively. Moreover, if each variable is coded with an identical sequence of numbers, it will be quite difficult to determine

Table 5.3 *Coding scheme for study of pseudo-titles and equivalent appositives*

Country	Sample number	Type	Correspondence relationship	Form	Length
US (1)	W2C001 (1)	Pseudo-title (1)	Total equivalence (1)	Simple NP (1)	One word (1)
Philippines (2)	W2C002 (2)	Appositive (2)	Determiner deletion (2)	Genitive NP (2)	Two words (2)
East Africa (3)	W2C003 (3)		Partial equivalence (3)	Multiple post-modification (3)	Three words (3)
Jamaica (4)	W2C004 (4)		N/A (4)		Four words (4)
New Zealand (5)	W2C005 (5)				Five words (5)
Great Britain (6)	etc.				Six or more words (6)
Singapore (7)					

Figure 5.1 PC-Tagger pop-up menu

whether a "2" assigned for Variable 1, for instance, is an error when the same numerical value is used for each of the other variables being studied. To reduce the incidence of errors, pseudo-titles and equivalent appositives were coded with a software program called PC Tagger (described in detail in Meyer and Tenney 1993). This program displays both the actual variables being studied in a particular corpus analysis, and the values associated with each variable (cf. figure 5.1).

Figure 5.1 contains a pseudo-title selected from ICE-Philippines, *MILF legal counsel Ferdausi Abbas*, and a pop-up window containing in the left column the "Tag Names" (i.e. variables) and in the right column "Tag Values." Because the "length" variable is selected, only the values for this variable appear. Since the pseudo-title is three words in length, the value "three" is selected. The program produces as output a separate file that converts all values to the numbers in table 5.3 associated with the values, and that can be exported to a statistical program. Because the analyst is working with the actual names of tags and their values, the results can be easily checked and corrected.

An alternative to using a program such as PC Tagger is to use a system containing codes that are mnemonic. For instance, the CHAT system (Codes for the Human Analysis of Transcripts), developed within the CHILDES Project (Child Language Data Exchange System; cf. section 1.3.8), contains numerous mnemonic symbols for coding differing kinds of linguistic data. To describe the morphological characteristics of the phrase *our family*, the codes "1P," "POSS," and "PRO" are used to describe *our* as a first-person-plural possessive pronoun; the codes "COLL" and "N" are used to describe *family* as a collective

noun (*CHAT Manual*: http://childes.psy.cmu.edu/pdf/chat.pdf). For the current study, a pseudo-title, for instance, could be coded as "PT" and a corresponding apposition as "AP." And even though a system such as this lacks numerical values, the mnemonic tags can be easily converted into numbers (e.g. all instances of "PT" can be searched for and replaced with the value "1").

5.3.3 Locating relevant constructions for a particular corpus analysis

The greatest amount of work in any corpus study will be devoted to locating the particular construction(s) being studied, and then assigning to these constructions the particular linguistic values being investigated in the study. In the past, when corpora were not available in computer-readable form, the analyst had to painstakingly extract grammatical information by hand, a very tedious process that involved reading through printed texts and manually recording grammatical information and copying examples by hand. Now that corpora exist in computer-readable form, it is possible to reduce the time it takes to conduct a corpus analysis by using software programs that can automate (to varying degrees) the extraction of grammatical information. In conducting a grammatical analysis of a corpus, the analyst can either learn a programming language such as Perl and then write "scripts" to extract the relevant grammatical information, or use a general-purpose software application, such as a concordancing program (described in detail below), that has been developed for use on any corpus and that can perform many common tasks (e.g. word searches) associated with any corpus analysis. Both approaches to corpus analysis have advantages and disadvantages, and neither approach will guarantee that the analyst retrieves precisely what is being sought: just about any corpus analysis will involve sorting through the data manually to eliminate unwanted constructions and to organize the data in a manner suitable to the analysis.

There exist a number of programming languages, such as Python, Visual Basic, or Perl, that can be powerful tools for analyzing corpora. Fernquest (2000) has written a number of Perl scripts that can, for instance, search a corpus and extract from it lines containing a specific phrase; that can calculate word frequencies, sorting words either alphabetically or by frequency of occurrence; and that can count "bigrams" in a text (i.e. two-word sequences) and organize the bigrams by frequency. Perl scripts are quite commonly available (cf., for instance, Melamud's 1996 file of Perl scripts), and can be modified to suit the specific grammatical analysis being conducted. Of course, much of what can be done with many Perl scripts can be more easily accomplished with, say, a good concordancing program. But as Sampson (1998) quite correctly observes, if those analyzing corpora have programming capabilities, they do not have to "rely on software produced by others which may not meet their needs."

If one is using someone else's scripts, using the scripts on a corpus is not that difficult. But writing new scripts requires programming skills that are probably

beyond the capabilities of the average corpus linguist. For these individuals, there exist a number of very useful software programs that require no programming skills and that can extract much useful information from corpora: the Linguistic Data Base (LDB) for analyzing the Nijmegen Corpus (cf. section 4.3); ICECUP for analyzing the British component of ICE (described below); Sara to analyze the British National Corpus (cf. Aston and Burnard 1998); and a variety of different concordancing programs available for use on PCs, Macintoshes, Unix work stations, even the World Wide Web (cf. appendix 2 for a listing of concordancing programs and also Hockey 2000: 49–65 for a survey of some of the more common programs). Of course, as is the case with using any new software program, the user will need to experiment with the program both to learn how to use it and to determine whether it can do what the user needs to accomplish in a given corpus analysis. Thus, before any corpus analysis is begun in earnest, the corpus analyst will want to experiment with more than one concordancing program to ensure that the best program is used for a given corpus analysis.

The ability of such programs to retrieve grammatical information depends crucially upon how easy it is to construct a search for a given grammatical construction. In a lexical corpus (a corpus containing raw text and no grammatical annotation), it is easiest to construct searches for particular "strings" of characters: a particular word or group of words, for instance, or classes of words containing specific prefixes or suffixes (e.g. the derivational morpheme *-un*, or a verb inflection such as *-ing*). In corpora that have been tagged or parsed, it is possible to search for particular tags and therefore retrieve the particular grammatical constructions (common nouns, adverbs, etc.) associated with the tags. However, even in tagged and parsed corpora, it may be difficult to automatically extract precisely what the analyst is studying. Therefore, just about any corpus study will at some stage involve the manual analysis of data. Pseudo-titles provide a good illustration of the strengths and limitations of the various tools that have been developed for corpus analysis.

A pseudo-title such as *linguist Noam Chomsky* consists of a common noun followed by a proper noun. Because nouns are open-class items, they are a very broad category. Consequently, it is highly likely that every pseudo-title discovered in a given corpus will contain a unique series of common and proper nouns. This characteristic of pseudo-titles makes it very difficult to use the most common corpus tool – the concordancing program – to automatically retrieve pseudo-titles.

The most common capability of any concordancing program is to conduct simple searches for words, groups of words, or words containing prefixes and suffixes; to display the results of a given search in KWIC (key word in context) format; and to calculate the frequency of the item being searched in the corpus in which it occurs. Figure 5.2 contains an example concordancing window displaying the results of a search for all instances of the conjunction *but* in the fully tagged and parsed British component of ICE (ICE-GB). This corpus

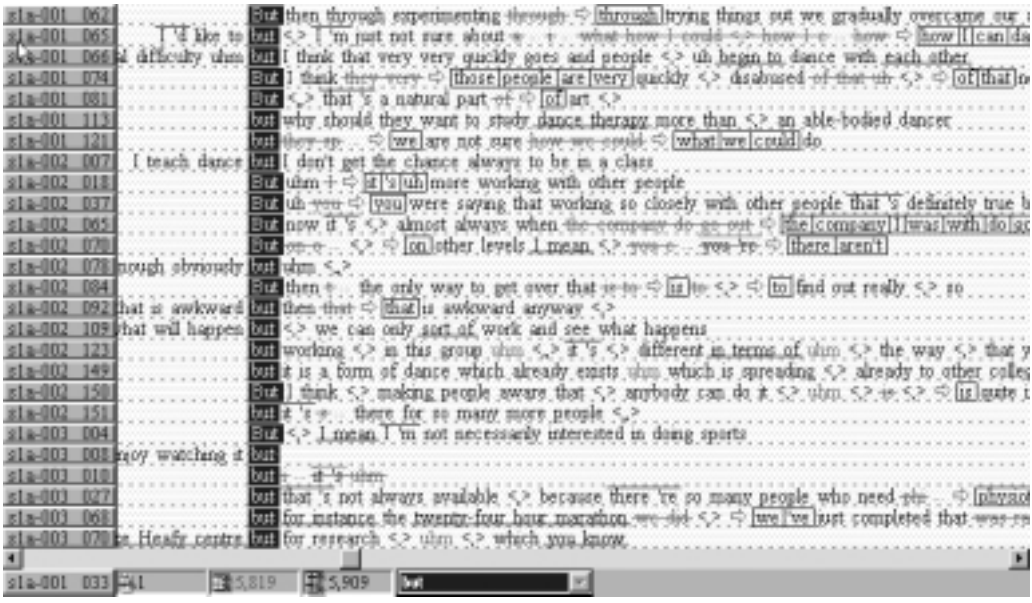


Figure 5.2 Concordancing window for the conjunction *but* in ICE-GB

can be searched by a text retrieval program, ICECUP, that has concordancing capabilities.

In figure 5.2, the conjunction *but* is displayed in KWIC format: all instances of *but* are vertically aligned for easy viewing, and displayed in the "text unit" (cf. section 4.1) in which they occur. In ICECUP (and most other concordancing programs) the context can be expanded to include, for instance, not just a single text unit but groups of text units occurring before and after the search item. At the bottom of the screen is a figure (5,909) specifying how many instances of *but* were found in the corpus.

While all concordancing programs can perform simple word searches, some have more sophisticated search capabilities. For instance, some programs can search not just for strings of characters but for "lemmas" (e.g. all the forms of the verb *be*) if a corpus has been "lemmatized," that is, if, for example, all instances of *be* (*am, was, are,* etc.) have been linked through the process of "lemmatization." In a corpus that has been lemmatized, if one wants information on the frequency of all forms of *be* in the corpus, it is necessary to simply search for *be*. In an unlemmatized corpus, one would have to search for every form of *be* separately.[5]

Many concordancing programs can also perform "wild card" searches. That is, a search can be constructed that finds not just strings of characters occurring

[5] Yasumasa Someya has produced a file of lemmatized verb forms for English that can be integrated into WordSmith and save the analyst the time of creating such a file manually. The file can be downloaded at http://www.liv.ac.uk/~ms2928/wordsmith/index.htm.

in an exact sequence but strings that may have other words intervening between them. For instance, a search of the correlative coordinator *not just . . . but*, as in *The movie was not just full of excessive violence but replete with foul language*, will have to be constructed so that it ignores the words that occur between *just* and *but*. To conduct such a search in a program such as Mono Conc Pro 2.0 (http://www.athel.com/mono.html#monopro), the number of words between the two parts of the search expression can be set, specifying that the search should find all instances of *not just . . . but* where, say, up to twenty words can occur between *just* and *but*. Wild card searches can additionally find parts of words: in WordSmith (http://www.liv.ac.uk/~ms2928/wordsmith/index.htm), searching for **ing* will yield all instances of words in a corpus ending in *-ing*. And wild card searches (as well as searches in general) can be extended to search for words tagged in a specific way in a given corpus, or a particular word tagged in a specific manner (e.g. all instances of *like* tagged as a conjunction).

Because pseudo-titles (and many other grammatical constructions) do not have unique lexical content, a simple lexical search will not retrieve any instances of pseudo-titles in a corpus. However, if pseudo-titles are studied in corpora that have been tagged or parsed, it is at least possible to narrow the range of structures generated in a search.

In a tagged or parsed corpus, it is possible to search not just for strings of characters but for actual grammatical categories, such as nouns, verbs, noun phrases, verb phrases, and so forth. Consequently, a tagged or parsed corpus will allow for a greater range of structures to be recovered in a given search. For instance, a concordancing program can be used to find possessive nouns by searching for the *strings 's* or *s'*. Although a search for these strings will certainly recover numerous possessive nouns, it will also retrieve many unwanted constructions, such as contractions (e.g. *John's leaving*). If possessives are tagged, however, it will be possible to search for the possessive tag, in a spoken or written text, and to recover only possessives, not extraneous constructions such as contractions. In ICE-GB, possessives are assigned the tag "genm" (genitive marker), and a search for all constructions containing this tag turned up all the instances of possessives in ICE-GB in a couple of seconds. Other programs, such as WordSmith or Mono Conc Pro 2.0, can also be used to search for tags.

Because in a tagged corpus only individual words are annotated, it can be difficult to construct searches for grammatical constructions having different part-of-speech configurations. Examples (8)–(10) below contain tagged instances of three pseudo-titles in ICE-GB:

(8) Community <**N(com,sing)**:1/3> liaison <**N(com,sing)**:2/3> officer <**N(com,sing)**:3/3> John <**N(prop,sing)**:1/2> Hambleton <**N(prop,sing)**:2/2>
(ICE-GB:W2C-009 #109:7)

(9) Liberal <**N(com,sing)**:1/3> Democrats <**N(com,sing)**:2/3> leader <**N(com,sing)**:3/3> Cllr <**N(prop,sing)**:1/3> John <**N(prop,sing)**:2/3> Hammond <**N(prop,sing)**:3/3>
(ICE-GB:W2C-009 #59:3)

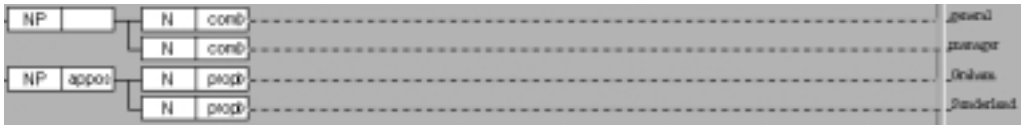


Figure 5.3 Parse tree for a sample pseudo-title in ICE-GB

(10) 59-year-old <**ADJ(ge)**> caretaker <**N(com,sing)**> Rupert <**N(prop,sing)**:1/2> Jones <**N(prop,sing)**:2/2> : <PUNC(col)> (ICE-GB:W2C-011 #67:2)

While the pseudo-titles in (8)–(10) consist of common and proper nouns, there are some notable differences: (8) and (10) contain two proper nouns, (9) three proper nouns; (10) begins with an adjective, (8) and (9) do not. There is thus no sequence of common nouns and proper nouns to be searched for that uniquely defines a pseudo-title. One could search for part of the pseudo-title that each of the constructions has in common (e.g. a proper noun preceding a common noun). But such a search is likely to turn up many structures other than pseudo-titles. When such a search was conducted on ICE-GB, a noun phrase followed by a vocative (11) was retrieved, as was an adverbial noun phrase followed by a proper noun (12):

(11) That's rubbish Aled (ICE-GB:S1A-068 #319:1:C)
(12) This time Hillier does find his man (ICE-GB:S2A-018 #120:1:A)

To retrieve particular grammatical constructions, it is therefore more desirable to use a parsed corpus rather than a tagged corpus because a parsed corpus will contain annotation describing higher-level grammatical constructions.

Of the various ICE corpora being used to study pseudo-titles, only ICE-GB has been parsed. To search ICE-GB for instances of pseudo-titles and corresponding appositives, it is first necessary to determine how such constructions have been parsed. Figure 5.3 contains a parse tree from ICE-GB for the pseudo-title *general manager Graham Sunderland*.

The pseudo-title in figure 5.3 has been parsed into two noun phrases, the first containing two common nouns and the second two proper nouns. The second noun phrase has the feature "appos," indicating that in ICE-GB pseudo-titles are considered a type of apposition.

Figure 5.4 contains a parse tree for what is considered an equivalent appositive structure, *Allen Chase, head of strategic exposure for NatWest Bank*.

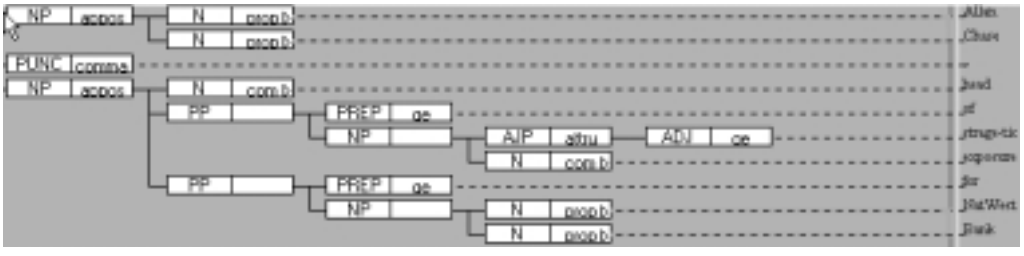


Figure 5.4 Parse tree for a sample appositive in ICE-GB

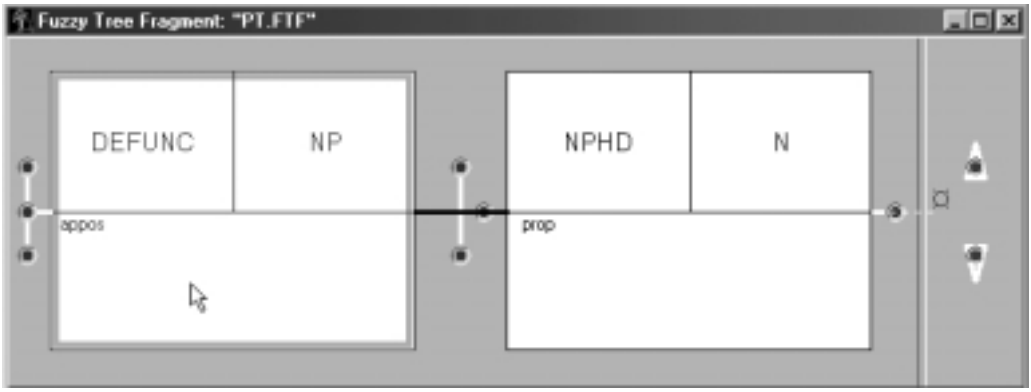


Figure 5.5 FTF for appositive proper nouns

Like the pseudo-title in figure 5.3, the construction in figure 5.4 has the form of two noun phrases containing various other structures. But unlike the pseudo-title in figure 5.3, the appositive in figure 5.4 contains the feature "appos" for both noun phrases, and additionally contains a parse marker for the mark of punctuation, the comma, that separates both noun phrases.

There are two ways to search for pseudo-titles and appositives in ICE-GB. Two separate searches can be constructed: one searching for two noun phrases with the feature "appos," and a second searching for two noun phrases with only the second one containing the feature "appos." Punctuation is irrelevant to both searches, as the search tool for ICE-GB, ICECUP, can be instructed to ignore punctuation. A more economical way to search for both pseudo-titles and equivalent appositives is to construct a search containing one structure that both constructions have in common: a single noun phrase containing proper nouns that have the feature "appos."

To find all instances of proper nouns annotated with the feature "appos," ICECUP requires that a "fuzzy tree fragment" (FTF) is constructed, 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 FTF. Figure 5.5 contains an FTF that will find all instances of proper nouns with the feature "appos."

This search was restricted to the press reportage section of ICE-GB, and in a matter of seconds, all of the relevant constructions were retrieved and displayed in KWIC format. Because the other components of ICE used in the study were not tagged or parsed, pseudo-titles in these components had to be identified manually, a process that took several days.

5.4 Subjecting the results of a corpus study to statistical analysis

After information is extracted from a corpus (or series of corpora), through either manual analysis or the use of a text retrieval program, it is

necessary to subject this information to some kind of statistical analysis. This analysis can be quite simple and involve no more than providing frequency counts of the particular constructions being investigated. However, if the goal of one's analysis is to determine whether similarities or differences exist in a corpus, it then becomes more necessary to apply specific statistical tests (such as chi-square) to determine whether the differences or similarities are statistically significant or not, that is, whether they are real or merely the result of chance.

Because many modern-day corpus linguists have been trained as linguists, not statisticians, it is not surprising that they have been reluctant to use statistics in their studies. Many corpus linguists come from a tradition that has provided them with ample background in linguistic theory and the techniques of linguistic description, but little experience in statistics. And as they begin doing analyses of corpora, they find themselves practicing their linguistic tradition within the realm of "numbers," the discipline of statistics, which many corpus linguists find foreign and intimidating. As a consequence, many corpus linguists have chosen not to do any statistical analysis of the studies they conduct, and to work instead mainly with frequency counts. For instance, the study of coordination ellipsis described in section 1.2 was based solely on frequency differences in the samples of speech and writing that were analyzed, and no statistical tests were applied to determine whether these differences were statistically different.

Recently, however, many corpus linguists have begun to take the role of statistics in corpus analysis much more seriously. A number of books have been written, such as Woods, Fletcher, and Hughes (1986), Kretzschmar and Schneider (1996), and Oakes (1998), that discuss statistical analysis of language data in considerable detail. Some text analysis programs, such as WordSmith, have statistical capabilities built into them so that as corpus linguists study particular linguistic constructions (e.g. collocations) they can perform statistical analyses to see whether the results they obtain are significant. And many of the commonly available statistical programs, such as SAS, SPSS, SYSTAT, or G, can now be run on PCs or Macintoshes and have very "friendly" user interfaces (cf. Kretzschmar 2000 for a review of the student version of SPSS, and a listing of other statistical packages that can be run on a PC or Macintosh).

If the corpus linguist finds any of these resources "intimidating," he or she can always consult a specialist in the area of statistics for advice. However, if this route is taken, it is important to realize that statistics is a vast and complex field of inquiry, and different academic disciplines will approach the use of statistical tests from differing perspectives: mathematicians, for instance, use statistics for different purposes than social scientists. Since corpus linguists are interested in studying linguistic variation within a corpus, it is most appropriate for them to follow the methodology of statistical analysis used in the social sciences.

However one obtains information on the use of statistics in corpus analysis, it is important to realize that by conducting a more rigorous statistical evaluation of their results, corpus linguists can not only be more confident about the results

they obtain but may even gain new insights into the linguistic issues they are investigating.

The statistical analysis of data is essentially a three-part process that involves:

1. evaluating the corpus from which the results are to be obtained to determine its suitability for statistical analysis;
2. running the appropriate statistical tests;
3. interpreting the results, and finding linguistic motivations for them.

Whether the latter part of step 3 – finding linguistic motivations for the results – is done first – in a series of hypothesis tested out in a corpus – or last depends upon the type of study one is doing, a point that will be discussed in greater detail in section 5.4.1. But following all of these steps is the best way to insure the validity and soundness of one's statistical analysis of a corpus.

5.4.1 Judging the suitability of a corpus for statistical analysis and determining the appropriate statistical tests to apply

Prior to conducting the study of pseudo-titles described in section 5.1, great care was taken in selecting a corpus that would yield valid results. Because pseudo-title usage has changed over time, texts were selected from a similar time-frame to ensure that the results reflected current usage trends. Even though only text fragments were used, a pilot study was conducted to ensure that sufficient numbers of pseudo-titles could be found in 2,000-word samples. Had texts from different time periods been mixed, or had the samples not turned up enough examples of pseudo-titles, it would have been difficult to trust the results that had been obtained, and considerable time would have been wasted conducting a study that for all practical purposes was invalid. The more the analyst is able to keep extraneous variables out of a corpus study, the more he or she will be able to trust the validity of the results obtained.

Many times, however, the analyst does not have the advantage of basing an analysis on a corpus that is methodologically "pure." This is particularly the case when using corpora developed in earlier eras, when those creating corpora did not have a "pilot corpus to guide their designs" (Biber 1993: 256) and the benefit of the years of experience that now exists in the design and creation of corpora. Also, the analyst may wish to do an analysis in an area where no good corpora are available. However, even if all or part of a corpus is not designed as ideally as the analyst would like, it is still possible to analyze the corpus and make generalizations based on the results that are obtained. Woods, Fletcher, and Hughes (1986: 55-6) advise that analysts "accept the results of each study, in the first place, as though any sampling had been carried out in a theoretically 'correct' fashion . . . [and] then look at the possibility that they may have been distorted by the way the sample was, in fact, obtained." If there are problems

with the sample, it is advisable "to attempt to foresee some of the objections that might be made about the quality of that material and either attempt to forestall criticism or admit openly to any serious defects."

Ideally, it would be most desirable to analyze only fully representative corpora so that it is not necessary to make the concessions that Woods, Fletcher, and Hughes (1986) advocate. But as long as the analyst is clear about the kind of corpus that was examined, open to the variables that might have affected the results, and sensitive to criticisms that may be leveled against the results, it is perfectly acceptable to base a linguistic analysis on a less than ideal corpus.

After determining whether the corpus being studied is suitable to the analysis being undertaken, the analyst needs to next consider what statistical tests are appropriate, and after applying the tests, whether any significant results obtained have some linguistic motivation: statistical significance is meaningless if it has no linguistic motivation. In deciding which statistical tests to apply, and how the results of such tests are best evaluated, it is useful to make the distinction that Biber (1988) makes between "macroscopic" and "microscopic" analyses. In "macroscopic" analyses, Biber (1988: 61) observes, the goal is to determine "the overall dimensions of variation in a language," that is, to investigate spoken vs. written English, or a variety of genres of English, and to determine which linguistic constructions define and distinguish these kinds of English. Macroscopic analyses are in a sense "inductive." In his analysis of linguistic variation in speech and writing, although Biber (1988: 65–75) pre-selected the corpora his analysis would be based on and the particular linguistic constructions he investigated, he did not initially go through the descriptive statistics that he generated, based on the occurrence of the constructions in the corpora, and attempt to determine which results looked meaningful and which did not before applying further statistical tests. Instead, he simply did a factor analysis of the results to determine which constructions tended to co-occur in specific genres and which did not. And it was not until he found that passives and nominalizations, for instance, tended to co-occur that he attempted to find a functional motivation for the results (Biber 1988: 80): that passives and nominalizations occurred together in genres (such as academic prose) in which abstract information predominated (Biber 1988: 119).

In "microscopic" analyses, on the other hand, the purpose is to undertake "a detailed description of the communicative functions of particular linguistic features": the analyst takes a particular grammatical construction, such as the relative clause, and attempts to study its function in various genres or linguistic contexts. Microscopic analyses are typically deductive rather than inductive: the analyst begins with a series of hypotheses and then proceeds to confirm or disconfirm them in the corpus being investigated. Of course, there will always be results that are unanticipated, but these results can be handled by proposing new hypotheses or reformulating old ones. Since the majority of corpus analyses involve microscopic rather than macroscopic analyses, the remainder of this chapter will focus on the various kinds of statistical tests that were applied to the results of the study of pseudo-titles detailed in section 5.3.

5.5 The statistical analysis of pseudo-titles

A good corpus study, as was argued in section 5.1, combines qualitative and quantitative research methods. In any corpus analysis, the balance between these methods will vary. This section explores this balance with respect to the various hypotheses proposed in section 5.2 concerning the use and structure of pseudo-titles, and demonstrates how these hypotheses can be confirmed or disconfirmed through simple "exploration" of the various samples of ICE included in the study as well as the application of various statistical tests.

5.5.1 Exploring a corpus

Previous research on pseudo-titles has documented their existence in American, British, and New Zealand press reportage, and demonstrated that because their usage was stigmatized, certain newspapers (particularly in the British press) prohibited their usage. To determine whether pseudo-titles have spread to the other varieties of English represented in ICE and whether their usage is stigmatized in these varieties, it is only necessary to examine examples from the various samples to see whether a given newspaper allows pseudo-title usage or not. And in simply examining examples in a corpus, one is likely to encounter "interesting" data. In the case of pseudo-titles, for instance, it was found that certain newspapers that prohibited pseudo-title usage did in fact contain pseudo-titles, a finding that reflects the well-known fact that practice does not always follow prescription.

Table 5.4 lists the number of newspapers containing or not containing pseudo-titles in the various regional components of ICE investigated. As table 5.4 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.

Table 5.4 *Number of newspapers containing pseudo-titles in various ICE components*

Country	Newspapers without pseudo-titles	Newspapers with 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%)

Although the results displayed in table 5.4 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 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; de Mönnink 1997): tests that ask individuals to identify constructions as acceptable or not, or that seek information from individuals about language usage. For pseudo-titles, one might give newspaper editors and writers a questionnaire that elicits their attitudes towards pseudo-titles, and asks them whether they consider certain borderline cases of pseudo-titles (such as *former President George Bush*; section cf. 5.3.1) to actually be pseudo-titles, and so forth.

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 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 contains 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.

5.5.2 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 (cf. section 5.3.2). 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 yielded the results displayed in table 5.5. Because so few examples of pseudo-titles and equivalent appositives were found in ICE-East Africa, ICE-Singapore, and ICE-Jamaica, the cross tabulations in table 5.5 (and elsewhere in this section) were restricted to the four ICE varieties (ICE-USA, ICE-Philippines, ICE-New Zealand, and ICE-Great Britain) from which a sufficient number of examples could be taken.

The results of the cross tabulation in table 5.5 yield raw numbers and percentages which suggest various trends. In ICE-USA, Phil, and 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,

Table 5.5 *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

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 5.5, 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 5.5 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 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

Table 5.6 *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

in the forty texts that were examined (cf. table 5.6), and the various statistical measurements that calculate whether a distribution is normal or not.

As the figures in table 5.6 indicate, the distribution of pseudo-titles across the forty 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 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. Figure 5.6 illustrates the "flat" and "skewed" nature of the curve. The horizontal axis plots the various number of pseudo-titles found in each of the forty texts, and the vertical axis the number of texts having a given frequency. The resultant curve is clearly not a bell curve.

Because most linguistic data behave the way that the data in table 5.5 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 5.5: 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

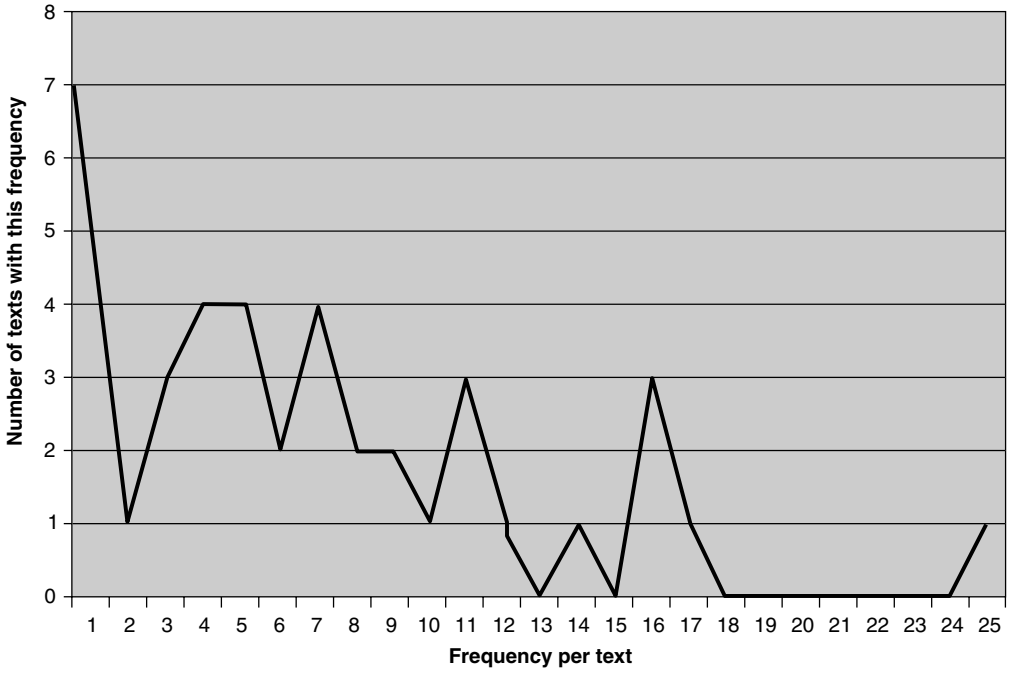


Figure 5.6 Pseudo-title frequency across samples

(i.e. the 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 5.7 presents the results of a chi-square analysis of the data in table 5.5.

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 5.5 yielded a value of 65.686. To interpret this number accurately, one first of all needs to know the "degrees of freedom" in a given dataset (i.e. the number of data points

Table 5.7 *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	$p \ll .000$

that may vary). Since table 5.5 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 5.7 is 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 5.5 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 from ICE-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 different chi-square tests will have to be performed:

.05	/	6	= .0083
Significance level		Number of tests performed	Corrected value

Table 5.8 contains the results of the comparison, from most significant differences down to least significant differences.

The results in table 5.8 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 constructions in ICE-GB is very different from their use in 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

Table 5.8 *Comparison of the distribution of pseudo-titles and corresponding appositives in individual ICE components*

Countries	Statistical test	Degrees of freedom	Value[6]	Significance level
NZ and GB	Chi-square	1	50.938	$p << 0.0001$
Phil and GB	Chi-square	1	44.511	$p << 0.0001$
US and GB	Chi-square	1	19.832	$p << 0.0001$
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.)

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 very useful 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 5.9 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 5.9. 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

Table 5.9 *Correspondence relationships for appositives in the samples from ICE components*

Country	Total equivalence	Determiner deletion	Partial equivalence	Total
USA	1 (2%)	14 (28%)	36 (71%)	51 (101%)
Phil	1 (2.6%)	8 (21%)	29 (76%)	38 (100%)
NZ	0 (0%)	13 (42%)	18 (58%)	31 (100%)
GB	8 (10%)	22 (28%)	48 (62%)	78 (100%)
Total	10 (5%)	57 (29%)	131 (66%)	198 (100%)

[6] In a 2×2 chi-square table, as Sigley (1997: 226) observes, the distribution is "binomial" rather than "continuous." It is therefore customary in a 2×2 table to use Yates' correction rather the normal Pearson chi-square value.

Table 5.10 *Correspondence relationships for appositives in the samples from ICE components (with combined cells)*

Country	Total equivalence/ determiner deletion	Partial equivalence	Total
USA	15 (29%)	36 (71%)	51 (100%)
Phil	9 (24%)	29 (76%)	38 (100%)
NZ	13 (42%)	18 (58%)	31 (100%)
GB	30 (39%)	48 (62%)	78 (101%)
Total	67 (34%)	131 (66%)	198 (100%)

Statistical test	Value	Degrees of freedom	Significance level
Chi-square	3.849	3	$p = .278$

of the chi-square statistic more valid. As was noted in section 5.1, one reason for recording the particular correspondence relationship 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 5.10 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

Table 5.11 *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 (101%)
NZ	66 (81%)	16 (20%)	82 (101%)
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	$p = .001$

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. Having a greater effect on the particular correspondence relation that was found was 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 had frequencies that were too low to apply the chi-square test. And combining categories, as is done in table 5.11, 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 phenomena." Dunning (1993: 62–3) notes that the chi-square test was designed to work with larger datasets that have 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 5.11, 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 5.11: 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

Table 5.12 *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

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.

(14) a. 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 Muslim Mindanao police chief Damming Unga (ICE-Philippines)

(15) a. 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)
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 5.12. Table 5.12 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" (one to four words or five words or more). Table 5.13 provides a cross tabulation of all of these variables.

Table 5.13 *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%)
Appos	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 chi-square analysis of the trends in table 5.13 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.

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 predetermined set of associations (i.e. only those associations that the analyst thinks exist in the data), or base the analysis on a "saturated model": 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:

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

Table 5.14 *Associations between various variables*

K	Degrees of freedom	Likelihood ratio	Probability	Probability
3	2	.155	.9254	.9300
2	7	488.010	.0000	.0000
1	11	825.593	.0000	.0000

Likelihood ratio and chi-square tests were conducted to determine whether there was a significant association between all three variables (16a), and between all possible combinations of two-way interactions (16b–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 5.14.

The first line in table 5.14 demonstrates that there were no associations between the three variables: the likelihood ratio score had probability 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 5.15 as being the strongest of all the associations tested. Interpreted in conjunction with the frequency distributions in table 5.13, the results in table 5.14 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

Table 5.15 *Strongest associations between variables*

	Degrees of freedom	Likelihood ratio	Probability
Type*form	2	246.965	.0000
Length*form	2	239.067	.0000

phrase will become. Therefore, pseudo-titles typically have forms (e.g. simple noun phrases) that that are short and non-complex structures, though as table 5.12 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 5.12 is very similar to the logistic regression models used in the Varbrul programs: IVARB (for MS-DOS) and GoldVarb (for the Macintosh) (cf. Sankoff 1987; Sigley 1997: 238–52). These programs have been widely used in sociolinguistics to test the interaction of linguistic 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 was a noun phrase with a noun or third-person pronoun as head.

Although the Varbrul programs have been used primarily in sociolinguistics to study the application of variable rules, Sigley (1997) demonstrates the value of the programs for corpus analysis as well in his study of relative clauses in the Wellington Corpus of New Zealand English. The advantage of the Varbul programs is that they were designed specifically for use in linguistic analysis and are thus easier to use than generic statistical packages, such as SPSS or SAS. But it is important to realize that these statistical packages, as the loglinear analysis in this section demonstrated, can replicate the kinds of statistical analyses done by the Varbrul programs. Moreover, as Oakes (1998: 1–51) demonstrates, these packages can perform a range of additional statistical analyses quite relevant to the concerns of corpus linguists, from those, such as the Pearson-Product Moment, that test correlations between variables to Regression tests, which test the effects that independent variables have on dependent variables.

5.6 Conclusions

To conduct a corpus analysis effectively, the analyst needs to plan out the analysis carefully. It is important, first of all, to begin the process with a very clear research question in mind, so that the analysis involves more than simply "counting" linguistic features. It is next necessary to select the appropriate

corpus for analysis: to make sure, for instance, that it contains the right kinds of texts for the analysis and that the corpus samples to be examined are lengthy enough. And if more than one corpus is to be compared, the corpora must be comparable, or the analysis will not be valid. After these preparations are made, the analyst must find the appropriate software tools to conduct the study, code the results, and then subject them to the appropriate statistical tests. If all of these steps are followed, the analyst can rest assured that the results obtained are valid and the generalizations that are made have a solid linguistic basis.

Study questions

1. What is the danger of beginning a corpus analysis without a clearly thought-out research question in mind?
2. How does the analyst determine whether a given corpus is appropriate for the corpus analysis to be conducted?
3. What kinds of analyses are most efficiently conducted with a concordancing program?
4. What kinds of information can be found in a tagged or parsed corpus that cannot be found in a lexical corpus?
5. The data in the table below are adapted from a similar table in Meyer (1996: 38) and contain frequencies for the distribution of phrasal (e.g. *John and Mary*) and clausal (e.g. *We went to the store and our friends bought some wine*) coordination in various samples of speech and writing from the International Corpus of English. Go to the web page given below at Georgetown University and use the "Web Chi-square Calculator" on the page to determine whether there is a difference between speech and writing in the distribution of phrasal and clausal coordination: http://www.georgetown.edu/cball/webtools/web_chi.html.

   Syntactic structures in speech and writing

Medium	Phrases	Clauses	Total
Speech	168 (37%)	289 (63%)	457 (100%)
Writing	530 (77%)	154 (23%)	684 (100%)

6 Future prospects in corpus linguistics

In describing the complexity of creating a corpus, Leech (1998: xvii) remarks that "a great deal of spadework has to be done before the research 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 a 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?

Planning a corpus. 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. We know, for instance, that what we plan to do with a corpus greatly determines how it is constructed: vocabulary studies necessitate larger corpora, grammatical studies (at least of relatively frequently occurring grammatical constructions) shorter corpora. The British National Corpus is the culmination of all the knowledge we have gained since the 1960s about what makes a good corpus.

But while it is of prime importance to descriptive corpus linguistics to create valid and representative corpora, in the field of natural language processing this is an issue of less concern. Obviously, the two fields have different interests: it does not require a balanced and representative corpus to train a parser or speech-recognition system. But it would greatly benefit the field of corpus linguistics if descriptive corpus linguists and more computationally oriented linguists and engineers worked together to create corpora. The British National Corpus is a good example of the kind of corpus that can be created when descriptive linguists, computational linguists, and the publishing industry cooperate. The TalkBank Project at Carnegie Mellon University and the University of Pennsylvania is a multi-disciplinary effort designed to organize varying interest groups engaged in the computational study of human and animal communication. One of the

interest groups, Linguistic Exploration, deals with the creation and annotation of corpora for purposes of linguistic research. The Michigan Corpus of Academic Spoken English (MICASE) is the result of a collaborative effort involving both linguists and digital librarians at the University of Michigan. Cross-disciplinary efforts such as these integrate the linguistic and computational expertise that exists among the various individuals creating corpora, they help increase the kinds and types of corpora that are created, and they make best use of the limited resources available for the creation of corpora.

Collecting and computerizing written texts. Because so many written texts are now available in computerized formats in easily accessible media, such as the World Wide Web, the collection and computerization of written texts has become much easier than in the past. It is no longer necessary for every written text to be typed in by hand or scanned with an optical scanner and then the scanning errors corrected. If texts are gathered from the World Wide Web, it is still necessary to strip them of html formatting codes. But this process can be automated with software that removes such markup from texts. Creating a corpus of written texts is now an easy and straightforward enterprise. The situation is more complicated for those creating corpora containing texts from earlier periods of English: early printed editions of books are difficult to scan optically with any degree of accuracy; manuscripts that are handwritten need to be manually retyped and the corpus creator must sometimes travel to the library where the manuscript is housed. This situation might be eased in coming years. There is an increased interest both among historical linguists and literary scholars in computerizing texts from earlier periods of English. As projects such as the Canterbury Project are completed, we may soon see an increased number of computerized texts from earlier periods.

Collecting and computerizing spoken texts. While it is now easier to prepare written texts for inclusion in a corpus, there is little hope for making the collection and transcription of spoken texts easier. For the foreseeable future, it will remain an arduous task to find people who are willing to be recorded, to make recordings, and to have the recordings transcribed. There are advantages to digitizing spoken samples and using specialized software to transcribe them, but still the transcriber has to listen to segments of speech over and over again to achieve an accurate transcription. Advances in speech recognition might automate the transcription of certain kinds of speech (e.g. speeches and perhaps broadcast news reports), but no software will be able to cope with the dysfluency of a casual conversation. Easing the creation of spoken corpora remains one of the great challenges in corpus linguistics, a challenge that will be with us for some time in the future.

Copyright restrictions. Obtaining the rights to use copyrighted material has been a perennial problem in corpus linguistics. The first release of the British National Corpus could not be obtained by anyone outside the European Union because of restrictions placed by copyright holders on the distribution of certain written texts. The BNC Sampler and second release of the entire corpus have

no distribution restrictions but only because the problematic texts are not included in these releases. The distribution of ARCHER has been on hold because it has not been possible to obtain copyright permission for many of the texts included in the corpus. As a result, access to the corpus is restricted to those who participated in the actual creation of the corpus – a method of distribution that does not violate copyright law. It is unlikely that this situation will ease in the future. While texts are more widely available in electronic form, particularly on the World Wide Web, getting permission to use these texts involves the same process as getting permission for printed texts, and current trends in the electronic world suggest that access to texts will be more restrictive in the future, not less. Therefore, the problem of copyright restrictions will continue to haunt corpus linguists for the foreseeable future.

Annotating texts with structural markup. The development of SGML-based annotation systems has been one of the great advances in corpus linguistics, standardizing the annotation of many features of corpora so that they can be unambiguously transferred from computer to computer. The Text Encoding System (TEI) has provided a system of corpus annotation that is both detailed and flexible, and the introduction of XML (the successor to HTML) to the field of corpus linguistics will eventually result in corpora that can be made available on the World Wide Web. There exist tools to help in the insertion of SGML-conformant markup to corpora, and it is likely that such tools will be improved in the future. Nevertheless, much of this annotation has to be inserted manually, requiring hours of work on the part of the corpus creator. We will have much better annotated corpora in the future, but it will still be a major effort to insert this annotation into texts.

Tagging and parsing. Tagging is now a standard part of corpus creation, and taggers are becoming increasingly accurate and easy to use. There will always be constructions that will be difficult to tag automatically and will require human intervention to correct, but tagged corpora should become more widespread in the future – we may even see the day when every corpus released has been tagged.

Parsing is improving too, but has a much lower accuracy than tagging. Therefore, much human intervention is required to correct a parsed text, and we have not yet reached the point that the team not responsible for designing the parser can use it effortlessly. One reason the British component of ICE (ICE-GB) took nearly ten years to complete was that considerable effort had to be expended correcting the output of the parsing of the corpus, particularly the spoken part. At some time in the future, parsing may be as routine as tagging, but because a parser has a much more difficult job than a tagger, we have some time to wait before parsed corpora will be widely available.

Text analysis. The most common text analysis program for corpora, the concordancer, has become an established fixture for the analysis of corpora. There are many such programs available for use on PCs, Macintoshes, and even the World Wide Web. Such programs are best for retrieving sequences of strings

(such as words), but many can now search for particular tags in a corpus, and if a corpus contains file header information, some concordancing programs can sort files so that the analyst can specify what he or she wishes to analyze in a given corpus: journalistic texts, for instance, but not other kinds of texts.

More sophisticated text analysis programs, such as ICECUP, are rare, and it is not likely that we will see as many programs of this nature in the future as concordancers. And a major problem with programs such as ICECUP and many concordancers is that they were designed to work on a specific corpus computerized in a specific format. Consequently, ICECUP works only on the British component of the International Corpus of English, and Sara on the BNC (though there are plans to extend the use of Sara to other corpora as well). The challenge is to systemize the design of corpora and concordancers so that any concordancer can work on any corpus. Of course, it is highly likely that the next generation of corpus linguists will have a much better background in programming. Thus, these corpus linguists will be able to use their knowledge of languages such as Perl or Visual Basic to write specific "scripts" to analyze texts, and as these scripts proliferate, they can be passed from person to person and perhaps make obsolete the need for specific text analysis programs to be designed.

Corpus linguistics has been one of the more exciting methodological developments in linguistics since the Chomskyan revolution of the 1950s. It reflects changing attitudes among many linguists as to what constitutes an adequate "empirical" study of language, and it has drawn upon recent developments in technology to make feasible the kinds of empirical analyses of language that corpus linguists wish to undertake. Of course, doing a corpus analysis will always involve work – more work than sitting in one's office or study and making up the data for a particular analysis – but doing a corpus analysis properly will always have its rewards and will help us advance the study of human language, an area of study that linguists of all persuasions would agree we still know relatively little about.

Appendix 1
Corpus resources

Cross references to resources listed in this table are indicated in boldface. The various resources are alphabetized by acronym or full name, depending upon which usage is most common.

The publisher has used its best endeavors to ensure that the URLs for external websites referred to in this book are correct and active at the time of going to press. However, the publisher has no responsibility for the websites and can make no guarantee that a site will remain live or that the content is or will remain appropriate.

Resource	*Description*	*Availability*
American National Corpus	Currently in progress; is intended to contain spoken and written texts that model as closely as possible the texts in the **British National Corpus**	Project website: http://www.cs.vassar.edu/~ide/anc/
American Publishing House for the Blind Corpus	25 million words of edited written American English	Originally created by IBM; described in Fillmore (1992)
ARCHER (A Representative Corpus of English Historical Registers)	1.7 million words consisting of various genres of British and American English covering the period 1650–1990	In-house corpus (due to copyright restrictions); an expanded version, ARCHER II, is underway
Bank of English Corpus	415 million words of speech and writing (as of October 2000); texts are continually added	Collins-Cobuild: http://titania.cobuild.collins.co.uk/boe_info.html
Bergen Corpus of London Teenage English (COLT)	500,000-word corpus of the speech of London teenagers from various boroughs; available online or as part of the British National Corpus	Project website: http://www.hd.uib.no/colt/
Birmingham Corpus	20 million words of written English	Evolved into **Bank of English Corpus**

British National Corpus (BNC)	100 million words of samples of varying length containing spoken (10 million words) and written (90 million words) British English	BNC website: http://info.ox.ac.uk/bnc/index.html
British National Corpus (BNC) Sampler	2 million words of speech and writing representing 184 samples taken from the British National Corpus	BNC website: http://info.ox.ac.uk/bnc/getting/sampler.html
Brown Corpus	One million words of edited written American English; created in 1961; divided into 2,000-word samples from various genres (e.g. press reportage, fiction, government documents)	See: **ICAME CD-ROM**
Cambridge International Corpus	100 million words of varying amounts of spoken and written British and American English, with additional texts being added continuously	CUP website: http://uk.cambridge.org/elt/reference/cic.htm
Cambridge Learners' Corpus	10 million words of student essay exams, with additional texts being added continuously	CUP website: http://uk.cambridge.org/elt/reference/clc.htm
Canterbury Project	Ultimate goal of the project is to make available in electronic form all versions of the *Canterbury Tales* and to provide an interface to enable, for instance, comparisons of the various versions	Project website: http://www.cta.dmu.ac.uk/projects/ctp/
Chemnitz Corpus	A parallel corpus of English and German translations	Project website: http://www.tu-chemnitz.de/phil/english/real/transcorpus/index.htm

Child Language Data Exchange System (CHILDES) Corpus	Large multi-lingual database of spoken language from children and adults engaged in first or second language acquisition	http://childes.psy.cmu.edu/ See also: MacWhinney (2000)
Corpora Discussion List (made available through the Norwegian Computing Centre for the Humanities)	Internet discussion list for issues related to corpus creation, analysis, tagging, parsing, etc.	http://www.hit.uib.no/corpora/welcome.txt
Corpus of Early English Correspondence	Two versions of English correspondence: the full version (2.7 million words) and a sampler version (450,000 words)	Project website: http://www.eng.helsinki.fi/doe/projects/ceec/corpus.htm Sampler version available on **ICAME CD-ROM**
Corpus of Middle English Prose and Verse	A large collection of Middle English texts available in electronic format	Project website: http://www.hti.umich.edu/c/cme/about.html
Corpus of Spoken Professional English	Approximately 2 million words taken from spoken transcripts of academic meetings and White House press conferences	Athelstan website: http://www.athel.com/cpsa.html
The Electronic *Beowulf*	A digital version of the Old English poem *Beowulf* that can be searched	Project website: http://www.uky.edu/~kiernan/eBeowulf/guide.htm
English–Norwegian Parallel Corpus	A parallel corpus of English and Norwegian translations: 30 samples of fiction and 20 samples of non-fiction in the original and in translation	Project website: http://www.hf.uio.no/iba/prosjekt/
The Expert Advisory Group on Language Engineering Standards (EAGLES)	Has developed "A Corpus Encoding Standard" containing guidelines for the creation of corpora	Project website: http://www.cs.vassar.edu/CES/

FLOB (Freiburg–Lancaster–Oslo–Bergen) Corpus	One million words of edited written British English published in 1991; divided into 2,000-word samples in varying genres intended to replicate the **LOB Corpus**	See: **ICAME CD-ROM**
FROWN (Freiburg–Brown) Corpus	One million words of edited written American English published in 1991; divided into 2,000-word samples in varying genres intended to replicate the **Brown Corpus**	See: **ICAME CD-ROM**
Helsinki Corpus	Approximately 1.5 million words of Old, Middle, and Early Modern English divided into samples of varying length	See: **ICAME CD-ROM**
Helsinki Corpus of Older Scots	Approximately 400,000 words of transcribed speech (recorded in the 1970s) from four rural dialects in England and Ireland	See: **ICAME CD-ROM**
Hong Kong University of Science and Technology Learner Corpus	25 million words of learner English written by first-year university students whose native language is Chinese	Contact: John Milton, Project Director, lcjohn@ust.hk
ICAME Bibliography	Extensive bibliography of corpus-based research created by Bengt Altenberg (Lund University, Sweden)	ICAME website: –1989: http://www.hd.uib.no/icame/icame-bib2.txt 1990–8: http://www.hd.uib.no/icame/icame-bib3.htm
ICAME CD-ROM	20 different corpora (e.g. Brown, LOB, Helsinki) in various computerized formats (DOS, Windows, Macintosh and Unix)	ICAME website: http://www.hit.uib.no/icame/cd/

International Corpus of English (ICE)	A variety of million-word corpora (600,000 words of speech, 400,000 words of writing) representing the various national varieties of English (e.g. American, British, Irish, Indian, etc.)	Three components now complete: Great Britain: http://www.ucl.ac.uk/english-usage/ice-gb/index.htm East Africa: http://www.tu-chemnitz.de/phil/english/real/eafrica/corpus.htm New Zealand: http://www.vuw.ac.nz/lals/corpora.htm#The New Zealand component of the International
ICECUP (ICE Corpus Utility Program)	Text retrieval software for use with ICE-GB	Survey of English Usage website: http://www.ucl.ac.uk/english-usage/ice-gb/icecup.htm
ICE-GB (British component of the International Corpus of English)	One million words of spoken and written British English fully tagged and parsed	See: **International Corpus of English**
International Corpus of Learner English (ICLE)	Approximately 2 million words of written English composed by non-native speakers of English from 14 different linguistic backgrounds	Project website: http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/introduction.html
Lampeter Corpus	Approximately 1.1 million words of Early Modern English tracts and pamphlets taken from various genres (e.g. religion, politics) from the period 1640–1740; contains complete texts, not text samples	Project website: http://www.tu-chemnitz.de/phil/english/real/lampeter/lamphome.htm

Lancaster Corpus	A precursor to the million-word **Lancaster–Oslo–Bergen (LOB) Corpus** of edited written British English	See: **Lancaster–Oslo–Bergen (LOB) Corpus**
Lancaster/IBM Spoken English Corpus	53,000 words of spoken British English (primarily radio broadcasts); available in various formats, including a parsed version	See: **ICAME CD-ROM**
Lancaster Parsed Corpus	A parsed corpus containing approximately 140,000 words from various genres in the **Lancaster–Oslo–Bergen (LOB) Corpus**	See: **ICAME CD-ROM**
Lancaster–Oslo–Bergen (LOB) Corpus	One million words of edited written British English published in 1961 and divided into 2,000-word samples; modeled after the **Brown Corpus**	See: **ICAME CD-ROM**
Linguistic Data Consortium (LDC)	For an annual fee, makes available to members a variety of spoken and written corpora of English and many other languages	LDC website: http://www.ldc.upenn.edu/
London Corpus	The original corpus of spoken and written British English first created in the 1960s by Randolph Quirk at the Survey of English Usage, University College London; sections of the spoken part are included in the **London–Lund Corpus**	Can be used on-site at the Survey of English Usage: http://www.ucl.ac.uk/english-usage/home.htm
London–Lund Corpus	Approximately 500,000 words of spoken British English from various genres (e.g. spontaneous dialogues, radio broadcasts) that has been prosodically transcribed	See: **ICAME CD-ROM**

Longman–Lancaster Corpus	A corpus available in orthographic form that contains approximately 30 million words of written English taken from various varieties of English world-wide	Longman website: http://www.longman-elt.com/dictionaries/corpus/lclonlan.html
Longman Learner's Corpus	10 million words of writing by individuals from around the world learning English as a second or foreign language	Longman website: http://www.longman-elt.com/dictionaries/corpus/lclearn.html
The Longman Spoken and Written English Corpus (LSWE)	Approximately 40 million words of samples of spoken and written British and American English	Described in Biber et al. (1999)
Map Task Corpus	Digitized transcriptions of individuals engaged in "task-oriented dialogues" in which one speaker helps another speaker replicate a route on a map	Project website: http://www.hcrc.ed.ac.uk/maptask/
Michigan Corpus of Academic Spoken English (MICASE)	Various types of spoken American English recorded in academic contexts: class lectures and discussions, tutorials, dissertation defenses	Searchable on the Web: http://www.hti.umich.edu/m/micase/
Nijmegen Corpus	A 130,000-word parsed corpus of written English	TOSCA Research website: http://lands.let.kun.nl/TSpublic/tosca/research.html
The Northern Ireland Transcribed Corpus of Speech	400,000 words of interviews with individuals speaking Hiberno-English from various regions of Northern Ireland	See Kirk (1992)
Penn–Helsinki Parsed Corpus of Middle English	1.3 million words of parsed Middle English taken from 55 samples found in the **Helsinki Corpus**	Project website: http://www.ling.upenn.edu/mideng/

Penn Treebank (Releases I and II)	A heterogeneous collection of speech and writing totaling approximately 4.9 million words; sections have been tagged and parsed	Linguistic Data Consortium (LDC) website: http://www.ldc.upenn.edu/Catalog/LDC95T7.html
Polytechnic of Wales Corpus	A 65,000-word parsed corpus of the speech of children ages 6–12 conversing in playgroups of three and in interviews with adults	See: **ICAME CD-ROM**
Santa Barbara Corpus of Spoken American English	Large corpus containing samples of varying length of different kinds of spoken American English: spontaneous dialogues, monologues, speeches, radio broadcasts, etc.	Project website: http://www.linguistics.ucsb.edu/research/sbcorpus/default.htm First release of corpus can be purchased from the Linguistic Data Consortium (LDC): http://www.ldc.upenn.edu/Catalog/LDC2000S85.html
Susanne Corpus	130,000 words of written English based on various genres in the Brown Corpus that have been parsed and marked up based on an "annotation scheme" developed for the project	Project website: http://www.cogs.susx.ac.uk/users/geoffs/RSue.html
Switchboard Corpus	2,400 telephone conversations between two speakers from various dialect regions in the United States; topics of conversations were suggested beforehand	Linguistic Data Consortium (LDC) website: http://www.ldc.upenn.edu/Catalog/LDC97S62.html
TalkBank Project	Cross-disciplinary effort to use computational tools to study human and animal communication	Project website: http://www.talkbank.org/

Tampere Corpus	A corpus proposed to consist of various kinds of scientific writing for specialized and non-specialized audiences	Described in Norri and Kytö (1996)
Text Encoding Initiative (TEI)	Has developed standards for the annotation of electronic documents	Project website: http://www.tei-c.org/
TIMIT Acoustic-Phonetic Continuous Speech Corpus	Various speakers from differing dialects of American English reading ten sentences containing phonetically varied sounds	Linguistic Data Consortium (LDC) website: http://www.ldc.upenn.edu/Catalog/LDC93S1.html
TIPSTER Corpus	Collection of various kinds of written English, such as *Wall Street Journal* and *Associated Press* news stories; intended for research in information retrieval	Linguistic Data Consortium (LDC) website: http://www.ldc.upenn.edu/Catalog/LDC93T3A.html
Wellington Corpus	One million words of written New Zealand English divided into genres that parallel the **Brown** and **LOB** corpora but that were collected between 1986 and 1990	See: **ICAME CD-ROM**
York Corpus	1.5 million words taken from sociolinguistic interviews with speakers of York English	See Tagliamonte (1998)

Appendix 2
Concordancing programs

PC/Macintosh-based programs

Conc (for the Macintosh)
John Thomson
available from Summer Institute of Linguistics
http://www.indiana.edu/~letrs/help-services/QuickGuides/about-conc.html

Concordancer for Windows
Zdenek Martinek in collaboration with Les Siegrist
http://www.ifs.tu-darmstadt.de/sprachlit/wconcord.htm

Corpus Presenter
Raymond Hickey
http://www.uni-essen.de/~lan300/corpus_presenter.htm

Corpus Wizard
Kobe Phoenix Lab, Japan
http://www2d.biglobe.ne.jp/~htakashi/software/CWNE.HTM

Lexa
Raymond Hickey
Available from Norwegian Computing Centre for the Humanities
http://www.hd.uib.no/lexainf.html

MonoConc Pro 2.0
Athelstan
http://www.athel.com/mono.html#monopro

ParaConc (for multilingual corpora)
Michael Barlow
http://www.athel.com/

Sara
British National Corpus
http://info.ox.ac.uk/bnc/sara/client.html

Tact
Centre for Computing in the Humanities, University of Toronto
http://www.chass.utoronto.ca:8080/cch/tact.html

WordCruncher (now called "Document Explorer")
Hamilton-Locke, Inc. (an older version is also available on the ICAME CD-ROM, 2nd edn.)
http://hamilton-locke.com/DocExplorer/Index.html

WordSmith
Mike Scott
Oxford University Press
http://www.oup.com/elt/global/catalogue/multimedia/wordsmithtools3/

Web-based programs

CobuildDirect
http://titania.cobuild.collins.co.uk/direct_info.html

KWiCFinder
http://miniappolis.com/KWiCFinder/KWiCFinderHome.html

The Michigan Corpus of Academic Spoken English (MICASE)
http://www.hti.umich.edu/micase/

Sara
Online version of the British National Corpus
http://sara.natcorp.ox.ac.uk

TACTWeb
http://kh.hd.uib.no/tactweb/homeorg.htm

References

Aarts, Bas (1992) *Small Clauses in English: The Nonverbal Types*. Berlin and New York: Mouton de Gruyter.

(2001) Corpus Linguistics, Chomsky, and Fuzzy Tree Fragments. In Mair and Hundt (2001). 5–13.

Aarts, Bas and Charles F. Meyer (eds.) (1995) *The Verb in Contemporary English*. Cambridge University Press.

Aarts, Jan and Willem Meijs (eds.) (1984) *Corpus Linguistics: Recent Developments in the Use of Computer Corpora*. Amsterdam: Rodopi.

Aarts, Jan, Pieter de Haan, and Nelleke Oostdijk (eds.) (1993) *English Language Corpora: Design, Analysis, and Exploitation*. Amsterdam: Rodopi.

Aarts, Jan, Hans van Halteren, and Nelleke Oostdijk (1996) The TOSCA Analysis System. In Koster and Oltmans (1996). 181–91.

Aijmer, Karin and Bengt Altenberg (eds.) (1991) *English Corpus Linguistics*. London: Longman.

Altenberg, Bengt and Marie Tapper (1998) The Use of Adverbial Connectors in Advanced Swedish Learners' Written English. In Granger (1998). 80–93.

Ammon, U., N. Dittmar, and K. J. Mattheier (eds.) (1987) *Sociolinguistics: An International Handbook of the Science of Language and Society*, vol. 2. Berlin: de Gruyter.

Aston, Guy and Lou Burnard (1998) *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Atwell, E., G. Demetriou, J. Hughes, A. Schiffrin, C. Souter, and S. Wilcock (2000) A Comparative Evaluation of Modern English Corpus Grammatical Annotation Schemes. *ICAME Journal* 24. 7–23.

Barlow, Michael (1999) MonoConc 1.5 and ParaConc. *International Journal of Corpus Linguistics* 4 (1). 319–27.

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.

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 (1994). 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 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.

Blachman, Edward, Charles F. Meyer, and Robert A. Morris (1996) The UMB Intelligent ICE Markup Assistant. In Greenbaum (1996a). 54–64.

Brill, Eric (1992) A Simple Rule-Based Part-of-Speech Tagger. *Proceedings of the 3rd Conference on Applied Natural Language Processing*. Trento: Italy.

Burnard, Lou (1995) The Text Encoding Initiative: An Overview. In Leech, Myers, and Thomas (1995). 69–81.

(1998) The Pizza Chef: A TEI Tag Set Selector. http://www.hcu.ox.ac.uk/TEI/pizza.html.

Burnard, Lou and C. M. Sperberg-McQueen (1995) TEI Lite: An Introduction to Text Encoding for Interchange. http://www.tei-c.org/Lite/index.html.

Burnard, Lou and Tony McEnery (eds.) (2000) *Rethinking Language Pedagogy from a Corpus Perspective*. Frankfurt: Peter Lang.

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 (1995). 54–61.

Chafe, Wallace, John Du Bois, and Sandra Thompson (1991) Towards a New Corpus of American English. In Aijmer and Altenberg (1991). 64–82.

Chomsky, Noam (1995) *The Minimalist Program*. Cambridge, MA: MIT Press.

Coates, Jennifer (1983) *The Semantics of the Modal Auxiliaries*. London: Croom Helm.

Collins, Peter (1991a) The Modals of Obligation and Necessity in Australian English. In Aijmer and Altenberg (1991). 145–65.

(1991b) *Cleft and Pseudo-Cleft Constructions in English*. Andover: Routledge.

Composition of the BNC. http://info.ox.ac.uk/bnc/what/balance.html.

Cook, Guy (1995) Theoretical Issues: Transcribing the Untranscribable. In Leech, Myers, and Thomas (1995). 35–53.

Corpus Encoding Standard (2000) http://www.cs.vassar.edu/CES/122.

Crowdy, Steve (1993) Spoken Corpus Design. *Literary and Linguistic Computing* 8. 259–65.

Curme, G. (1947) *English Grammar*. New York: Harper and Row.

Davies, Mark (2001) Creating and Using Multi-million Word Corpora from Web-based Newspapers. In Simpson and Swales (2001). 58–75.

de Haan, Pieter (1984) Problem-Oriented Tagging of English Corpus Data. In Aarts and Meijs (1984). 123–39.

Du Bois, John, Stephan Schuetze-Coburn, Susanna Cumming, and Danae Paolino (1993) Outline of Discourse Transcription. In Edwards and Lampert (1993). 45–89.

Dunning, Ted (1993) Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19 (1). 61–74.

Eckman, Fred (ed.) (1977) *Current Themes in Linguistics*. Washington, DC: John Wiley.
Edwards, Jane (1993) Principles and Contrasting Systems of Discourse Transcription. In Edwards and Lampert (1993). 3–31.
Edwards, Jane and Martin Lampert (eds.) (1993) *Talking Data*. Hillside, NJ: Lawrence Erlbaum.
Ehlich, Konrad (1993) HIAT: A Transcription System for Discourse Data. In Edwards and Lampert (1993). 123–48.
Elsness, J. (1997) *The Perfect and the Preterite in Contemporary and Earlier English*. Berlin and New York: Mouton de Gruyter.
Fang, Alex (1996) AUTASYS: Automatic Tagging and Cross-Tagset Mapping. In Greenbaum (1996a). 110–24.
Fernquest, Jon (2000) Corpus Mining: Perl Scripts and Code Snippets. http://www.codearchive.com/home/jon/program.html.
Fillmore, Charles (1992) Corpus Linguistics or Computer-Aided Armchair Linguistics. In Svartvik (1992). 35–60.
Finegan, Edward and Douglas Biber (1995) *That* and Zero Complementisers in Late Modern English: Exploring ARCHER from 1650–1990. In Aarts and Meyer (1995). 241–57.
Francis, W. Nelson (1979) A Tagged Corpus – Problems and Prospects. In Greenbaum, Leech, and Svartvik (1979). 192–209.
(1992) Language Corpora B.C. In Svartvik (1992). 17–32.
Francis, W. Nelson and H. Kučera (1982) *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin.
Fries, Udo, Gunnel Tottie, and Peter Schneider (eds.) (1994) *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, 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.
Garside, Roger, Geoffrey Leech, and Anthony McEnery (eds.) (1997) *Corpus Annotation*. London: Longman.
Garside, Roger and Nicholas Smith (1997) A Hybrid Grammatical Tagger: CLAWS 4. In Garside, Leech, and McEnery (1997). 102–121.
Gavioli, Laura (1997) Exploring Texts through the Concordancer: Guiding the Learner. In Anne Wichmann, Steven Fligelstone, Tony McEnery, and Gerry Knowles (eds.) (1997) *Teaching and Language Corpora*. London: Longman. 83–99.
Gillard, Patrick and Adam Gadsby (1998) Using a Learners' Corpus in Compiling ELT Dictionaries. In Granger (1998). 159–71.
Granger, Sylvianne (1993) International Corpus of Learner English. In Aarts, de Haan, and Oostdijk (1993). 57–71.
(1998) *Learner English on Computer*. London: Longman.
Greenbaum, Sidney (1973) Informant Elicitation of Data on Syntactic Variation. *Lingua* 31. 201–12.
(1975) Syntactic Frequency and Acceptability. *Lingua* 40. 99–113.
(1984) Corpus Analysis and Elicitation Tests. In Aarts and Meijs (1984). 195–201.
(1992) A New Corpus of English: ICE. In Svartvik (1992). 171–79.

(ed.) (1996a) *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press.

(1996b) *The Oxford English Grammar*. Oxford: Oxford University Press.

Greenbaum, Sidney, Geoffrey Leech, and Jan Svartvik (eds.) (1979) *Studies in English Linguistics*. London: Longman.

Greenbaum, S. and Meyer, C. (1982) Ellipsis and Coordination: Norms and Preferences. *Language and Communication* 2:137–49.

Greenbaum, Sidney and Jan Svartvik (1990) The London–Lund Corpus of Spoken English. In Svartvik (1990). 11–45.

Greenbaum, Sidney and Ni Yibin (1996) About the ICE Tagset. In Greenbaum (1996a). 92–109.

Greenbaum, Sidney, Gerald Nelson, and Michael Weizman (1996) Complement Clauses in English. In Thomas and Short (1996). 76–91.

Greene, B. B. and G. M. Rubin (1971) Automatic Grammatical Tagging. Technical Report. Department of Linguistics: Brown University.

Hadley, Gregory (1997) Sensing the Winds of Change: An Introduction to Data-Driven Learning. http://web.bham.ac.uk/johnstf/winds.htm.

Haegeman, Lilliane (1987) Register Variation in English: Some Theoretical Observations. *Journal of English Linguistics* 20 (2). 230–48.

(1991) *Introduction to Government and Binding Theory*. Oxford: Blackwell.

Halteren, Hans van and Theo van den Heuvel (1990) *Linguistic Exploitation of Syntactic Databases. The Use of the Nijmegen Linguistic DataBase Program.* Amsterdam: Rodopi.

Haselrud, V. and Anna-Brita Stenström (1995) Colt: Mark-up and Trends. *Hermes* 13. 55–70.

Hasselgård, Hilde (1997) Sentence Openings in English and Norwegian. In Ljung (1997). 3–20.

Hickey, Raymond, Merja Kytö, Ian Lancashire, and Matti Rissanen (eds.) (1997) *Tracing the Trail of Time. Proceedings from the Second Diachronic Corpora Workshop.* Amsterdam: Rodopi.

Hockey, Susan (2000) *Electronic Texts in the Humanities*. Oxford: Oxford University Press.

Järvinen, Timo (1994) Annotating 200 Million Words: The Bank of English Project. Proceedings of COLING '94, Kyoto, Japan. http://www.lingsoft.fi/doc/engcg/Bank-of-English.html.

Jespersen, Otto (1909–49) *A Modern English Grammar on Historical Principles*. Copenhagen: Munksgaard.

Johansson, Stig and Knut Hofland (1994) Towards an English–Norwegian Parallel Corpus. In Fries, Tottie, and Schneider (1994). 25–37.

Johansson, Stig and Jarle Ebeling (1996) Exploring the English–Norwegian Parallel Corpus. In Percy, Meyer, and Lancashire (1996).

Johns, Tim F. (1994) From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-driven Learning. In Odlin (1994). 293–313.

Kalton, Graham (1983) *Introduction to Survey Sampling*. Beverly Hills, CA: Sage.

Kennedy, Graeme (1996) Over *Once* Lightly. In Percy, Meyer, and Lancashire (1996). 253–62.

Kettemann, Bernhard (1995) On the Use of Concordancing in ELT. *TELL&CALL* 4. 4–15.

Kirk, John (1992) The Northern Ireland Transcribed Corpus of Speech. In Leitner (1992). 65–73.

Koster, C. and E. Oltmans (eds.) (1996) *Proceedings of the first AGFL Workshop.* Nijmegen: CSI.

Kretzschmar, William A., Jr. (2000) Review of *SPSS Student Version 9.0 for Windows. Journal of English Linguistics* 28 (3). 311–13.

Kretzschmar, William A., Jr. and E. Schneider (1996) *Introduction to Quantitative Analysis of Linguistic Survey Data*. Los Angeles: Sage.

Kretzschmar, William A., Jr., Charles F. Meyer, and Dominique Ingegneri (1997) Uses of Inferential Statistics in Corpus Linguistics. In Ljung (1997). 167–77.

Kytö, M. (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 edn. Department of English: University of Helsinki.

Labov, W. (1972) The Transformation of Experience in Narrative Syntax. In *Language in the Inner City*. Philadelphia: University of Pennsylvania Press. 354–96.

Landau, Sidney (1984) *Dictionaries: The Art and Craft of Lexicography*. New York: Charles Scribner.

Leech, Geoffrey (1992) Corpora and Theories of Linguistic Performance. In Svartvik (1992). 105–22.

(1997) Grammatical Tagging. In Garside, Leech, and McEnery (1997). 19–33.

(1998) Preface. In Granger (1998). 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 (1997). 34–52.

Leitner, Gerhard (ed.) (1992) *New Directions in English Language Corpora*. Berlin: Mouton de Gruyter.

León, Fernando Sánchez and Amalio F. Nieto Serrano (1997) Retargeting a Tagger. In Garside, Leech, and McEnery (eds.). 151–65.

Ljung, Magnus (ed.) (1997) *Corpus-based Studies in English*. Amsterdam: Rodopi.

MacWhinney, Brian (1996) The CHILDES System. *American Journal of Speech-Language Pathology* 5. 5–14.

(2000) *The CHILDES Project: Tools for Analyzing Talk*. 3rd edn., vol. 1: *Transcription Format and Programs*, vol 2: *The Database*. Mahwah, NJ: Erlbaum.

Mair, Christian (1990) *Infinitival Complement Clauses in English*. Cambridge University Press.

(1995) Changing Patterns of Complementation, and Concomitant Grammaticalisation, of the Verb *Help* in Present-Day British English. In Aarts and Meyer (1995). 258–72.

Mair, Christian and Marianne Hundt (eds.) (2001) *Corpus Linguistics and Linguistic Theory*. Amsterdam: Rodopi.

Maniez, François (2000) Corpus of English Proverbs and Set Phrases. Message posted on the Corpora List, 24 January. http://www.hit.uib.no/corpora/2000–1/0057.html.

Marcus, M., B. Santorini, and M. 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 Ljung (1997). 211–26.

Melčuk, Igor A. (1987) *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press.

Melamed, Dan (1996) 170 General Text Processing Tools (Mostly in PERL5). http://www.cis.upenn.edu/~melamed/genproc.html.

Meurman-Solin, Anneli (1995) A New Tool: The Helsinki Corpus of Older Scots (1450–1700). *ICAME Journal* 19. 49–62.

Meyer, Charles F. (1992) *Apposition in Contemporary English*. Cambridge University Press.

(1995) Coordination Ellipsis in Spoken and Written American English. *Language Sciences* 17. 241–69.

(1996) Coordinate Structures in the British and American Components of the International Corpus of English. *World Englishes* 15. 29–41.

(1997) Minimal Markup for ICE Texts. *ICE NEWSLETTER* 25. http://www.cs.umb.edu/~meyer/icenews2.html.

(1998) Studying Usage on the World Wide Web. http://www.cs.umb.edu/~meyer/usage.html.

Meyer, Charles F. and Richard Tenney (1993) Tagger: An Interactive Tagging Program. In Souter and Atwell (1993). 302–12.

Meyer, Charles F., Edward Blachman, and Robert A. Morris (1994) Can You See Whose Speech Is Overlapping? *Visible Language* 28 (2). 110–33.

Milton, John and Robert Freeman (1996) Lexical Variation in the Writing of Chinese Learners of English. In Percy, Meyer, and Lancashire (1996). 121–31.

Mindt, Dieter (1995) *An Empirical Grammar of the English Verb*. Berlin: Cornelsen.

Mönnink, Inga de (1997) Using Corpus and Experimental Data: A Multimethod Approach. In Ljung (1997). 227–44.

Murray, Thomas E. and Carmen Ross-Murray (1992) On the Legality and Ethics of Surreptitious Recording. *Publication of the American Dialect Society* 76. 15–75.

(1996) Under Cover of Law: More on the Legality of Surreptitious Recording. *Publication of the American Dialect Society* 79. 1–82.

Nelson, Gerald (1996) Markup Systems. In Greenbaum (1996a). 36–53.

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.

Newmeyer, Frederick (1998) *Language Form and Language Function*. Cambridge, MA: MIT Press.

Nguyen, Long, Spyros Matsoukas, Jason Devenport, Daben Liu, Jay Billa, Francis Kubala, and John Makhoul (1999) Further Advances in Transcription of Broadcast News. *Proceedings of the 6th European Conference on Speech Communication and Technology*, vol. 2. Edited by G. Olaszy, G. Nemeth, K. Erdohegy (aka EuroSpeech '99). European Speech Communication Association (ESCA). 667–70.

Norri, Juhani and Merja Kytö (1996) A Corpus of English for Specific Purposes: Work in Progress at the University of Tampere. In Percy, Meyer, and Lancashire (1996). 159–69.

Oakes, Michael P. (1998) *Statistics for Corpus Linguistics*. Edinburgh University Press.

Odlin, Terrence (ed.) (1994) *Perspectives on Pedagogical Grammar*. New York: Cambridge University Press.

Ooi, Vincent (1998) *Computer Corpus Lexicography*. Edinburgh University Press.

Oostdijk, Nelleke (1991) *Corpus Linguistics and the Automatic Analysis of English*. Amsterdam: Rodopi.

Pahta, Päivi and Saara Nevanlinna (1997) Re-phrasing in Early English. Expository Apposition with an Explicit Marker from 1350 to 1710. In Rissanen, Kytö, and Heikkonen (1997). 121–83.

Percy, Carol, Charles F. Meyer, and Ian Lancashire (eds.) (1996) *Synchronic Corpus Linguistics*. Amsterdam: Rodopi.

Porter, Nick and Akiva Quinn (1996) Developing the ICE Corpus Utility Program. In Greenbaum (1996a). 79–91.

Powell, Christina and Rita Simpson (2001) Collaboration between Corpus Linguists and Digital Librarians for the MICASE Web Search Interface. In Simpson and Swales (2001). 32–47.

Prescott, Andrew (1997) The Electronic Beowulf and Digital Restoration. *Literary and Linguistic Computing* 12. 185–95.

Quinn, Akiva and Nick Porter (1996) Annotation Tools. In Greenbaum (1996a). 65–78.

Quirk, Randolph (1992) On Corpus Principles and Design. In Svartvik (1992). 457–69.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik (1972) *A Grammar of Contemporary English*. London: Longman.

(1985) *A Comprehensive Grammar of the English Language*. London: Longman.

Renouf, Antoinette (1987) Corpus Development. In Sinclair (1987). 1–40.

Rissanen, Matti (1992) The Diachronic Corpus as a Window to the History of English. In Svartvik (1992). 185–205.

(2000) The World of English Historical Corpora: from Cædmon to the Computer Age. *Journal of English Linguistics* 28 (1). 7–20.

Rissanen, Matti, Merja Kytö, and Kirsi Heikkonen (eds.) (1997) *English in Transition: Corpus-based Studies in English Linguistics and Genre Styles*. *Topics in English Linguistics* 23. Berlin and New York: Mouton de Gruyter.

Robinson, Peter (1998) New Methods of Editing, Exploring, and Reading *The Canterbury Tales.* http://www.cta.dmu.ac.uk/projects/ctp/desc2.html.

Rocha, Marco (1997) A Probabilistic Approach to Anaphora Resolution in Dialogues in English. In Ljung (1997). 261–79.

Rydén, Mats (1975) Noun-Name Collocations in British Newspaper Language. *Studia Neophilologica* 67. 14–39.

Sampson, Geoffrey (1998) Corpus Linguistics User Needs. Message posted to the Corpora List, 29 July. http://www.hd.uib.no/corpora/1998–3/0030.html.

Samuelsson, Christer and Atro Voutilainen (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*. Madrid: Association for Computational Linguistics. 246–53.

Sánchez, Aquilino and Pascual Cantos (1997) Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish. *International Journal of Corpus Linguistics* 2. 259–80.

Sanders, Gerald (1977) A Functional Typology of Elliptical Coordinations. In Eckman (1977). 241–70.

Sankoff, David (1987). Variable rules. In Ammon, Dittmar, and Mattheier (1987). 984–97.

Schmied, Josef (1996) Second-Language Corpora. In Greenbaum (1996a). 182–96.

Schmied, Josef and Hildegard Schäffler (1996) Approaching Translationese Through Parallel and Translation Corpora. In Percy, Meyer and Lancashire (1996). 41–55.

Schmied, Josef, and Claudia Claridge (1997) Classifying Text- or Genre-Variation in the *Lampeter Corpus of Early Modern English Texts*. In Hickey, Kytö, Lancashire, and Rissanen (1997). 119–35.

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 Burnard and McEnery (2000). 43–9.

Simpson, Rita and John Swales (eds.) (2001) *Corpus Linguistics in North America*. Ann Arbor: University of Michigan Press.

Sinclair, John (ed.) (1987) *Looking Up: An Account of the COBUILD Project*. London: Collins.

(1991) *Corpus, Concordance, Collocation*. Oxford University Press.

(1992) Introduction. *BBC English Dictionary*. London: HarperCollins. x–xiii.

Smith, Nicholas (1997) Improving a Tagger. In Garside, Leech and McEnery (1997). 137–50.

Souter, Clive and Eric Atwell (eds.) (1993) *Corpus-Based Computational Linguistics*. Amsterdam: Rodopi.

Sperberg-McQueen, C. M. and Lou Burnard (eds.) (1994a) *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. http://etext.virginia.edu/TEI.html.

(1994b) A Gentle Introduction to SGML. In *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. http://etext.lib.virginia.edu/bin/tei-tocs?div=DIV1&id=SG.

Stenström, Anna-Brita and Gisle Andersen (1996) More Trends in Teenage Talk: A Corpus-Based Investigation of the Discourse Items *cos* and *innit*. In Percy, Meyer, and Lancashire (1996). 189–203.

Svartvik, J. (ed.) (1990) *The London–Lund Corpus of Spoken English*. 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 University 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, D. (1989) *Talking Voices: Repetition, Dialogue, and Imagery in Conversational Discourse*. Cambridge University Press.

Tapanainen, Pasi and Timo Järvinen (1997) A Non-projective Dependency Parser. http://www.conexor.fi/anlp97/anlp97.html [also published in *Procs. ANLP-97*. ACL. Washington, DC].

The Independent Style Manual, 2nd edn. (1988) London: *The Independent*.

The Spoken Component of the BNC. http://info.ox.ac.uk/bnc/what/spok_design.html.

The Written Component of the BNC. http://info.ox.ac.uk/bnc/what/writ_design.html.

Thomas, Jenny and Mick Short (eds.) (1996) *Using Corpora for Language Research*. London: Longman.

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 (1995). 168–80.

Tottie, G. (1991) *Negation in English Speech and Writing. A Study in Variation*. San Diego: Academic Press.

van Halteren, Hans and Theo van den Heuvel (1990) *Linguistic Exploitation of Syntactic Databases*. Amsterdam: Rodopi.

Voutilainen, Atro (1999) A Short History of Tagging. In Hans van Halteren (ed.) *Syntactic Wordclass Tagging* (1999). Dordrecht: Kluwer.

Voutilainen, Atro and Mikko Silvonen (1996) A Short Introduction to ENGCG. http://www.lingsoft.fi/doc/engcg/intro.

Wheatley, B., G. Doddington, C. Hemphill, J. Godfrey, E.C. Holliman, J. McDaniel, and D. Fisher (1992) Robust Automatic Time Alignment of Orthographic Transcriptions with Unconstrained Speech. *Proceedings of ICASSP-92*, vol. 1. 533–6.

Willis, Tim (1996) Analysing the Lancaster/IBM Spoken English Corpus (SEC) Using the TOSCA Analysis System (for ICE): Some Impressions from a User. In Percy, Meyer, and Lancashire (1996). 237–51.

Wilson, A. and Tony McEnery (1994) *Teaching and Language Corpora*. Technical Report. Department of Modern English Language and Linguistics, University of Lancaster.

Wilson, Andrew and Jenny Thomas (1997) Semantic Annotation. In Garside, Leech and McEnery (1997). 53–65.

Woods, A, P. Fletcher, and A. Hughes (1986) *Statistics in Language Studies.* Cambridge University Press.

Index