English Corpus Linguistics
An Introduction

CHARLES F. MEYER
University of Massachusetts at Boston

CAMBRIDGE
UNIVERSITY PRESS

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

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

http://www.cambridge.org

First published 2002

Printed in the United Kingdom at the University Press, Cambridge

Typefaces Times New Roman 10/13 pt. and Formata *System* LaTeX 2$_\varepsilon$ [TB]

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

Library of Congress Cataloguing in Publication data

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

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

Contents

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 elipt 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