The Handbook of English Linguistics
Edited by Bas Aarts, April McMahon

3 English Corpus Linguistics

TONY MCENERY AND
COSTAS GABRIELATOS

1 Introduction

Since the 1960s, electronic corpora have come to prominence as a resource used by linguists. While their use remains a source of debate and controversy to this day (see for example Newmeyer 2003; Prodromou 1997; Seidlehofer 2003: 77–123; Widdowson 1991) their contribution to linguistics in general, and English linguistics in particular, as well as to language teaching, is now widely acknowledged. Corpus tools have not only strengthened the position of descriptive linguistics, but have also enhanced theoretically oriented linguistic research. This contribution has been felt most strongly in English linguistics, as it was pioneering work undertaken on English language corpora, such as the *Brown* corpus (Francis and Kučera 1964), which set the agenda for much of the work that has been undertaken using corpora since then. In this chapter we will examine the nature of corpus linguistics, review the general contribution of corpora to linguistic theory and then explore in more depth the contribution of corpora in four major areas:

- language description in general, and the production of reference resources in particular;
- lexicogrammar and the lexical approach to language analysis and description (lexical grammar);
- the teaching of English as a foreign language;
- the study of language change, with particular reference to the role that corpora have to play in theoretically informed accounts of language change.

2 The Nature of Corpus Linguistics

Introductory books on corpus linguistics are generally at pains to assert that corpus linguistics is not a branch of linguistics, nor a linguistic theory, but a methodology, one of the possible ways of 'doing' linguistics (e.g. Biber et al. 1998: 3–4; Kennedy 1998: 7; McEnery and Wilson 2001: 2; Meyer 2002: xi). For some, the term *corpus linguistics* is now synonymous with *empirical linguistics* (e.g. Sampson 2001: 6). However, it has been argued that, although it is not a linguistic theory in itself, corpus linguistics is more than just a methodology; rather, it is "a new research enterprise" and "a new philosophical approach to the subject" (Leech 1991: 106). It has also been proposed that corpus linguistics has a "theoretical status" (Tognini-Bonelli 2001: 1), in that observations of language facts lead to the formulation of hypotheses and generalizations, which are then unified in a theoretical statement; corpora need not simply be used to test existing theories, particularly ones formulated mainly on the basis of intuitions (2001: 2; see also section 3 below).

A point that all writers defining corpus linguistics agree upon is that corpus linguistics is empirical, in that it examines, and draws conclusions from, attested language use, rather than intuitions. This is not to say that intuitions play no role in corpus linguistics, but that they do not provide the data for analysis, nor do intuitions supersede the empirical evidence. Also, as a rule, corpus linguistics examines samples, however large, of language use, as it is typically impossible to capture the entirety of a language in a corpus. Yet corpus linguistics can examine entireties if, for example, the corpus content is limited in terms of one or more of the following: authorship, topic, and place and date of publication. For example, it is feasible to build corpora containing the entire work of a novelist, or the text of a newspaper over a period of time.

Another central characteristic of modern corpus linguistics is the use of computers; in fact, the term 'corpus linguistics' is now synonymous with 'computer corpus linguistics' (e.g. Leech 1992: 106). Hunston (2002: 20) makes explicit the dual function of computers in facilitating the collection and storage of large amounts of language data, and in enabling the development of the software that is used to access and analyze the corpus data. The pivotal role of computers in corpus linguistics is such that corpus linguistics has also been defined as a branch of computational linguistics (e.g. Oostdijk 1991: 2). The benefits of the use of computers in corpus linguistics are substantial. Computers and software programs have enabled researchers to collect, store and manage vast amounts of data relatively quickly and inexpensively. Data analysis and processing is fast and, in many instances, automated. The use of computers "gives us the ability to *comprehend*, and to account for, the contents of . . . corpora in a way which was not dreamed of in the pre-computational era of corpus linguistics" (Leech 1992: 106). Automated processes also allow for the replicability of studies, and the checking of the statistical reliability of results.

Although corpus linguistics does not downplay the importance of the qualitative interpretation of the data (e.g. Mair 1991), it does, nevertheless, have a strong focus on quantitative information, that is, frequency counts and statistical measures. The absolute and relative frequency of linguistic items features heavily in most, if not all, corpus studies. Statistical information based on the frequency of occurrence of language items is at the heart of probabilistic accounts of language (e.g. Halliday 1991). Statistical measures on the strength of lexical co-occurrence, which also take into account the relative frequency of the co-occurring items, play a central role in much of the research done within the neo-Firthian paradigm (e.g. Stubbs 2002).

3 Debates in Corpus Linguistics

It was mentioned in the previous section that corpus linguistics is viewed primarily as a methodology, not a theory. However, this should not be understood to imply that corpus linguistics is theory-free. The focus and method of research, as well as the type of corpus selected for a study, is influenced by the theoretical orientation of the researchers, explicit or implicit. Kennedy's statement that corpus linguistics has "a tendency sometimes to focus on lexis and lexical grammar rather than pure syntax" (1998: 8) is a case in point. Methodologically, corpus linguistics is equally diverse and encompasses different approaches to corpus building and use.[1] The main points of tension in corpus linguistics, which are interconnected, concern the relation between theory and data, the utility of corpus annotation,[2] and the role of intuitions.

These tensions have been formalized in the distinction between *corpus-based* and *corpus-driven* approaches to linguistics (e.g. Tognini-Bonelli 2001). This distinction is not acknowledged by all corpus linguists, and it has been felt by some to be overstated (Aarts 2002: 121), because "the worlds of the corpus-based and of the corpus-driven linguist may not be all that far apart as they are made out to be" (p. 123). However, since at the centre of this distinction lie the issues outlined above, the definitions of the corpus-based and corpus-driven approaches can serve as a springboard for the discussion of these issues.

In the corpus-based approach, the corpus is mainly used to "expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study" (Tognini-Bonelli 2001: 65). Although the intuitive basis of the theories being tested is seen as a weakness of this approach, it is not as much the target of criticism as the attitudes to, or techniques for, dealing with discrepancies between theoretical statements and corpus data that are supposed to characterize corpus-based linguists. Corpus annotation is a central feature of three techniques that are used. The first is to "insulate the data,"[3] that is, either to dismiss data that do not fit the theory, or to make the data fit the theory, for example, by annotating the corpus according to the theory (2001: 68–71). The second technique is to reduce the data to

"a set of orderly categories which are tractable within existing descriptive systems" (2001: 68), again by annotating the corpus. The criticism here is two-pronged: the annotation scheme is based on a pre-conceived theory, and the manual annotation of the training corpus is influenced by both the theory and the annotator's intuitions. The third technique is "building the data into a system of abstract possibilities, a set of paradigmatic choices available at any point in the text" (2001: 74), and is strongly associated with Halliday's probabilistic view of grammar (e.g. 1991, 1992). This stance is criticized mainly on two related grounds: its focus is predominantly paradigmatic rather than syntagmatic, that is, it is concerned with grammar rather than lexis (Tognini-Bonelli 2001: 75–7), and, consequently, requires an annotated corpus, since "grammatical patterns . . . are not easily retrievable from a corpus unless it is annotated" (2001: 77).

The basic tenet of the corpus-driven approach is that any "theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus" (2001: 84). Corpus-driven research aims at discovering facts about language free from the influence of existing theoretical frameworks, which are considered to be based on intuitions, and, therefore, are not comprehensive or reliable. Consequently, research is carried out on unannotated corpora, as annotation would impose a restrictive theoretical taxonomy on the data. A further characteristic of this approach is that it makes no distinction between lexis and grammar, as that, too, would require using existing distinctions, which may not be supported by the corpus data. Finally, in the corpus-driven approach the starting point of research is the patterning of orthographic words.[4] The remainder of this section will discuss views on the role of theory, intuitions, and annotation. As these issues are interrelated, their discussion will overlap to some extent.

As far as the role of theory in corpus linguistic research is concerned, it is more helpful to regard different approaches as falling between two end-points of a continuum, rather than belonging to one of two polar extremes. At one end, the corpus is used to find evidence for or against a given theory, or one or more theoretical frameworks are taken for granted;[5] at the other, the observed patterns in the corpus data are used as a basis from which to derive insights about language, independent of pre-existing theories and frameworks, with a view to developing a purely empirical theory. Of course this distinction begs the question of whether data observation and analysis can ever be atheoretical. It is interesting to note that the corpus-based approach, which is criticized in Tognini-Bonelli (2001), is associated with corpus research influenced by the work of Leech (e.g. 1991) or Halliday (e.g. 1991), and is presented as typically prioritizing "the information yielded by syntactic rather than lexical patterns" (Tognini-Bonelli 2001: 81), whereas the corpus-driven approach, which is proposed in Tognini-Bonelli (2001) , is associated with corpus research influenced by the work of Sinclair (e.g. 1991) and Firth's contextual theory of meaning, and favors a focus on lexical patterning. This indicates that the distinction is not only methodological, but also theoretical. Hunston and Francis (2000: 250), who have located their study of *pattern grammar* within the corpus-driven

paradigm, state that their method "is indeed theory-driven," as "theories are, in a sense, constructed by methods."

Our view is that an atheoretical approach is not possible and hence the idea of corpus-driven approaches to language must be seen as an idealized extreme, because, as Stubbs (1996: 47) notes, "the concept of data-driven linguistics must confront the classic problem that there is no such thing as pure induction . . . The linguist always approaches data with hypotheses and hunches, however vague." Sampson (2001: 124) shifts the focus from the formulation of hypotheses to their testing:

> We do not care how a scientist dreams up the hypotheses he puts forward in the attempt to account for the facts – he will usually need to use imagination in formulating hypotheses, they will not emerge mechanically from scanning the data. What is crucial is that any hypothesis which is challenged should be *tested* against interpersonally observable, objective data.

The testing of hypotheses on corpus data is related to the use of intuitions and the annotation of corpora. Sinclair (2004a: 39) contrasts two attitudes in corpus linguistics research in a manner which reveals that, for those working within the corpus-driven paradigm, the use of annotation is seen as interconnected with the use of intuition:

> Some corpus linguists prefer to research using plain text, while others first prepare the texts by adding various analytic annotations. The former group express reservations about the reliability of intuitive "data," whereas the latter group, if obliged, will reject corpus evidence in favor of their intuitive responses.

One explanation for this connection is that adherence to a given theory is expected to have influenced the linguist to such an extent that the categories and structures recognized by the theory have become part of his/her intuitions. Sampson (2001: 135) highlights the role of schooling in the forming of intuitions: "Certainly we have opinions about language before we start doing linguistics . . . In some cases our pre-scientific opinions about language come from what we are taught in English lessons, or lessons on other languages, at school." Similarly, Sinclair (2004a: 40) sees intuition not as a "gut reaction to events, [but] educated in various ways, and sophisticated." It can be argued that the influence of education on intuitions about language is more pronounced in linguists whose education and training involves familiarization with a number of theories, and, not uncommonly, in-depth study of a specific theoretical framework.

Although the usefulness of intuitions in the forming of hypotheses has been challenged by corpus-driven linguists, there seems to be a consensus that intuitions are unavoidable in the interpretation of corpus data (e.g. Hunston 2002: 65). However, Sinclair (2004a: 47) has argued that there is a way for "keeping . . . intuition temporarily at bay." The technique seems to involve the decontextualization of the observed patterns and a temporary disassociation of form and meaning, and is aided by examining the vertical patterns of the

key word in a concordance, or slotting in alternative words in a frame (e.g. *on the __ of*). Sinclair (2004a: 47–8) argues that:

> Since the essence of finding the meaning-creating mechanisms in corpora is the comparison of the patterns – as physical objects and quasi-linguistic units – with the meanings, it is valuable to be able at times to study one without the other. This takes a little skill and practice, but to my mind should be an essential part of the training of a corpus linguist.

One criticism of annotation is that it imposes the categories of a theoretical framework on the data, a practice which may interfere with finding evidence against the theory, or with discovering language features that the theory does not predict. There is also disagreement on whether annotation adds information, and therefore "value," to the corpus (Leech 1997a: 2), or whether it "loses information" (Sinclair 2004a: 52), because it assigns only one unalterable tag, when the word may not clearly belong to one existing category. Finally, reservations have been expressed regarding the degree to which corpus researchers are aware of the theoretical assumptions underlying different annotation schemes (e.g. Hunston 2002: 67; Sinclair 2004a: 55–6).

Leech (1997a: 6–8) outlines three "practical guidelines, or standards of good practice" (p. 6) for the annotation of corpora, and three further "maxims [applicable] both to the compilers and users of annotated corpora" (pp. 6–7), which partly address these reservations.

1 The raw corpus should be recoverable.
2 The annotation should be extricable.
3 The corpus user should have access to documentation providing information about the annotation scheme, the rationale behind it, the annotators, the place of annotation, and comments on the quality of annotation.
4 The annotation scheme "does not come with any 'gold standard' guarantee, but is offered as a matter of practical usefulness only" (p. 6).
5 The annotation scheme should be "based as far as possible on **consensual** or theory-neutral analyses of the data" (p. 7) [boldface in original].
6 "No one annotation scheme should claim authority as an absolute standard" (p. 7).

There is agreement on the necessity for the unannotated version of a corpus to be available to researchers (Leech 1997a: 6; Sinclair 2004a: 50–1). There also seems to be an area of consensus on the need for researchers to be aware of the theoretical principles behind the annotation scheme. Although Leech's point (3) above does not include the explicit statement of the theory informing the annotation, it can be argued that the theoretical framework should be inferable from the information given in the documentation.

The main point of concern, that of the imposition of a theory on the data, seems to be largely unresolved. Linguists of the corpus-driven persuasion would consider existing annotation schemes to be influenced by intuition-based

theories, and, therefore, restricting. Proponents of annotation would see the annotated corpus as "a repository of linguistic information, because the information which was implicit in the plain text has been made explicit through concrete annotation" (McEnery and Wilson 2001: 32). However, some consensus, albeit implicit, regarding the categories used in annotation schemes seems to exist, as corpus-driven studies do make use of what might be called traditional categories, such as 'verb,' 'preposition,' 'object,' 'clause' and 'passive,' without a definition (e.g. Hunston and Francis 2000; Tognini-Bonelli 2001), which indicates that they are treated as given. Furthermore, if, as Sinclair (2004a: 47–8) proposes, it is feasible for linguists to distance themselves from their intuitions, it can be argued that it is also feasible to adopt an informed and critical approach towards the annotation. Finally, irrespective of the perceived usefulness of the annotated corpus as a product, the annotation process can reveal the strengths and limitations of the theory informing the annotation scheme and lead to its modification – a process which is consistent with an empirical approach. Aarts (2002: 122) argues that "the only way to test the correctness and coverage of an existing description is to formalize it into an annotation system and test it on a corpus . . . It is the annotation *process*, rather than its *result* (i.e. an annotated corpus) that matters" (see also Leech 1992: 112).

Although within the corpus-driven paradigm annotation is seen as counter-productive when the corpus is used for theoretically oriented research, it is deemed acceptable when the corpus is annotated with a view to being used in an "application" (Sinclair 2004a: 50–6), that is, "the use of language tools in order to achieve a result that is relevant outside the world of linguistics . . . [such as] a machine that will hold a telephone conversation, or a translating machine or even a dictionary" (p. 55). An argument that can be advanced on the basis of this view is that if applications relying on a corpus which has been annotated according to a theoretical framework are successful, then this can be regarded as an indication that the theory affords helpful insights into actual language use.

Undoubtedly, there are pitfalls and limitations in uncritically using an annotated corpus. However, the use of an unannotated corpus has its own pitfalls and limitations. An unannotated electronic corpus lends itself to the examination of forms and their patterns, as the software exists that will produce a concordance of a word-form for manual examination, or statistical measures of the strength of its collocation patterns, from an unannotated corpus.[6] However, an unannotated corpus is of little, if any, use if the research focus is upon grammatical categories, semantic notions or pragmatic functions. Tognini-Bonelli (2001: 89–90) concedes that "while collocation is instantly identifiable on the vertical axis of an alphabetical concordance, colligation represents a step in abstraction and is therefore less immediately recognizable unless the text is tagged with precisely the required grammatical information." Sampson (2001: 107) agrees that, "in general, more complex forms of investigation may only be possible if the computer has access to some form of detailed linguistic analysis of the text."

Also, the interpretation of concordance lines (e.g. Hunston 2002: 38–66), that is, the manual examination of concordances in order to identify patterns, which is a frequently used technique of corpus-driven linguists, is open to what we might call 'implicit annotation.' That is, while examining concordance lines, researchers may assign grammatical or semantic roles to words or configurations of words, either unwittingly, influenced by tradition or their education, or consciously, refraining from using established roles and patterns.

What becomes evident from the discussion of tensions in corpus linguistics is that theoretical and methodological issues are interconnected. Therefore, these issues will, inevitably, be revisited in the remainder of this chapter. In sum, when considered from specific theoretical or methodological viewpoints, different approaches to corpus linguistics appear to have merits, as well as problems and limitations. However, when considered from the viewpoint of linguistics in general, the current diversity in corpus research can only be seen as an indication of health, and should be welcomed. The next section examines in some detail the theoretical assumptions and methodological positions of what has been termed the lexical approach (or lexical grammar), and which lies behind the corpus-driven approach to linguistic research.

4 Lexicogrammar and Lexical Grammar

A major contribution to English corpus linguistics is the body of work related to lexicogrammar. This work will be covered at some length in this chapter, both because it has a salience in corpus linguistics and because it undoubtedly represents a unique contribution made by corpus linguists to linguistic theory. The idea of lexicogrammar stems from the tension caused by generating a strict distinction between the lexical and grammatical, and distinguishing between the syntagmatic and paradigmatic dimensions of language, what Sinclair (1991: 109–10) also terms the *slot and filler model* or the *open choice principle*. In this view of language, words are combined according to grammatical principles, that is, grammatical structures have grammatically defined 'slots' that can be filled by any semantically appropriate word fulfilling the grammatical criteria. This view was challenged by Firth through the concept of collocation, which concerns "syntagmatic relations between words as such, not between categories" (Stubbs 1996: 35). Sinclair (1991: 110) notes that "the open-choice principle does not provide substantial enough restraints on consecutive words. We would not produce normal text simply by operating the open-choice principle." It is unsurprising then that researchers have focused on breaking down this distinction.

Following Firth, Sinclair (1991) proposed the *idiom principle* to account for syntagmatic relations between words which cannot be explained in terms of grammar. His approach, which he terms *lexical grammar* (Sinclair 2004b: 164), is to discover generalizations about language by examining the interaction and patterning of lexis. In a sense, this entails approaching grammar via lexis,

as evidence from large corpora "suggests that grammatical generalizations do not rest on a rigid foundation, but are the accumulation of the patterns of hundreds of individual words and phrases" (1991: 100; see also Halliday 1992: 64).

Halliday (1991, 1992) presents lexis and grammar as being "the same thing seen by different observers" (1992: 63) or "complementary perspectives" (1991: 32), and prefers the term *lexicogrammar*. He presents lexis and grammar as two ends of a continuum, with grammar being the "deeper" end. As examples of the 'lexis' end he cites sense relations, for example the different types of associations between the word *run* and the words *walk*, *hop*, and *jog* respectively. Examples of 'grammar' are polarity, mood, and transitivity, whereas prepositions and systems of modality occupy a middle position. In Halliday's words, one should keep in mind that "if you interrogate the system grammatically you will get grammar-like answers and if you interrogate it lexically you get lexis-like answers" (1992: 64).

At the root of this approach to lexical meaning and language description is Firth's notion of *meaning by collocation* (Firth 1951/1957). The notion of collocation, and its application to defining lexical meaning, as well as its use as the basis for a lexical description of English by a group of linguists often dubbed 'Neo-Firthians,' has had a profound influence, not only on the scope and focus of research in English linguistics, but also on the compilation of corpora and the use of corpus-based methodologies.

Firth (1951/1957: 194–6) introduced the term *collocation* to refer to one of the three "levels" of meaning he distinguished: "meaning by collocation," the "conceptual or idea approach to the meaning of words" and "contextual meaning." Later, Halliday (1966) and Sinclair (1966) took this idea further and, without abandoning collocation as defining meaning, introduced the notion that patterns of collocation can form the basis for a lexical analysis of language alternative to, and independent of, a grammatical analysis. In fact, they regarded the two levels of analysis as being complementary, with neither of the two being subsumed by the other. However, it is interesting to note that Halliday's and Sinclair's approaches take as their respective starting points the two ends of the lexicogrammar continuum. Halliday 'interrogates' language grammatically, aiming "to build the dictionary out of the grammar" (1992: 63); Sinclair 'interrogates' language lexically seeking to discover "facts . . . that cannot be got by grammatical analysis" (1966: 410). In fact, Sinclair (2004b: 164) distinguishes *lexicogrammar* from *lexical grammar*: "[lexicogrammar] is fundamentally grammar with a certain amount of attention to lexical patterns within the grammatical frameworks; it is not in any sense an attempt to build together a grammar and lexis on an equal basis."

But perhaps the contrast of lexis and grammar obscures more than it illuminates, as "lexical items do not contrast with each other in the same sense as grammatical classes contrast" (Sinclair 1966: 411). Stubbs explains that collocation, which is at the centre of a lexical description of language, is "a purely lexical relation, non-directional and probabilistic, which ignores any syntactic

relation between the words" (2001: 64). Sinclair sees the lexical item "balanc[ing] syntagmatic and paradigmatic patterns, using the same descriptive categories to describe both dimensions" (1998: 23).

Firth (1968) defined *collocation* as a relation between words, and introduced the notion of *colligation* for relations at the grammatical level, that is the interrelation of "word and sentence classes or of similar categories" and not "between words as such" (1968). Sinclair (1991: 170) defined collocation as "the occurrence of two or more words within a short space of each other in a text," and proposed a span of four to five words on either side of the node (i.e. the word whose collocations are examined) (1968: 105–6, 121). From early on, collocation was understood in relation to the probability that two words will co-occur. Firth (1968: 181) stated that collocation is "an order of mutual expectancy" (see also Hoey 1991: 7; Sinclair 1966: 418; Stubbs 2001: 64). *Collocation* is now understood in a somehow less restricted sense than that defined by Firth (1968: 181), and may include the co-occurrence of lexis and grammatical categories (Stubbs 2001: 112), and in some cases it is understood as only the latter, that is "the grammatical company a word keeps" (Hoey 1997: 8). A third relation between words, also a feature of the idiom principle (Sinclair 1991: 110), is that of *semantic prosody*, defined as the "consistent aura of meaning with which a form is imbued by its collocates" (Louw 1993: 157; see also Sinclair 1991: 112). Stubbs (2001: 111–12) makes a finer distinction between *semantic preference*, the "relation between a lemma or word-form and a set of semantically related words," and *discourse prosody* (or *semantic prosody*), "a feature which extends over more than one unit in a linear string . . . Since they are evaluative, prosodies often express the speaker's reason for making the utterance, and therefore identify functional discourse units."

Underlying the lexical approach to the analysis of language are a series of requirements relating to research methodology. Firstly, and perhaps most importantly, the researcher's reliance on intuitions and traditional concepts and categories has to be minimized as much as possible, if it cannot be excluded altogether (see Phillips 1989: 5; Stubbs 1996: 22). Sinclair (1991: 39) sees a role for intuitions "in evaluating evidence rather than creating it." Consequently, as we saw in section 3 above, corpus annotation is treated with caution, if not viewed unfavorably, as it represents the imposition of categories and preconceptions on the part of the annotator or the programmer of the annotation software.[7] However, there are researchers within the Neo-Firthian paradigm who have adopted the view that, despite its perceived problems and limitations, there are cases when annotation may be acceptable (e.g. Hunston 2002: 80–94), or even desirable (e.g. Teubert 1999).

Intuitions can be sidestepped if the linguist consciously tries to suppress his/her intuitions when examining concordances (Sinclair 2004a: 47; see also section 3), or if the analysis is automatic and relies on the computation of statistical results (Sinclair 1966: 413). This notion has been taken as far as treating a text as "essentially a statistical phenomenon" (Phillips 1989: 17). The only phase of lexical research where intuitions, or rather "intentionality and

human reasoning," are considered acceptable, or at least unavoidable, is in "the validation and interpretation of [the] processed data" (Teubert 1999; see also Firth 1957: 1, 29; Stubbs 1996: 47).

Secondly, any generalizations at the lexical level must be informed by the collocational patterns of lexical items (Sinclair 1991: 8). What is taken as the unit of corpus-based analysis is the orthographic word, rather than the lemma, as *a priori* lemmatization is seen as introducing the analyst's subjective intuitions (p. 41).[8] A further reason is that different forms of what is traditionally considered the same 'word' (i.e. lemma) have been observed to display different patterns. Sinclair (1991: 53–65, 154–6) examined the senses and syntactic patterns of the different forms of the lemma YIELD in a 7.3 million word corpus of written texts, and provided evidence of correlation between the different forms of YIELD, on the one hand, and meaning and syntactic patterns on the other.

Thirdly, corpora should contain whole texts, not samples. In Firthian and neo-Firthian linguistics, language is seen as a social phenomenon, which is observable in discourse and text. This consideration, coupled with findings indicating that different parts of a text demonstrate different patterning in terms of lexical and grammatical frequencies and relations (cf. Stubbs 1996: 32–4), points towards the building of corpora that contain whole texts rather than samples (e.g. Sinclair 1991: 19).

Finally, this approach favors very large corpora.[9] Since the basis of the analysis is words rather than categories, the researcher needs to examine a large number of instances of specific word-forms in order to be able to recognize patterns. The problem for lexis-based research is that the smaller the corpus, the higher the percentage of *hapax legomena*, that is, words which occur only once (e.g. Kennedy 1998: 100; Sinclair 1991: 18–19), or words with too low a frequency for dependable generalizations to be made. Table 3.1 summarizes the main characteristics of lexis-based language description and their implications for corpus building and corpus-based research methodology.

It may not be an overstatement to say that the main impetus, if not the driving force, behind much English corpus-based lexical research is the development of a description of language which takes as its basic units lexical items, rather than grammatical categories, such as noun or verb (Stubbs 1996: 35). In fact, some corpus linguists (e.g. Teubert 1999) have gone as far as to effectively equate corpus linguistics with collocation-based research on lexical description. This view is disputable, but what seems to be indisputable is that this approach to language description has prompted the use of lexis-based research methodologies, particularly those examining collocations and the behavior of different forms of the same root, by studies that do not intend to contribute to the lexis-based paradigm of linguistic research. For example, given the ease of exploring corpora lexically[10] and Halliday's views on language, many English corpus linguists have started to view language lexically in order to get 'lexis-like' answers to what have been traditionally treated as grammatical questions. The following section provides examples of a range of methodological and theoretical approaches to corpus linguistics.

Table 3.1 Characteristics of lexis-based approaches

Characteristics of lexis-based language description	Implications for corpus building and analysis
• Analysis needs to be as free as possible from introspective assumptions.	• No annotation. • No lemmatization: calculation of the collocations of orthographic words. • Reliance on (automatic) statistical analysis.
• Linguistic features do not normally show the same distribution across different sections of a text.	• Corpora should contain whole texts, not samples.
• A large number of occurrences of different words is needed for dependable analysis.	• Corpora should be as large as possible.

5 Corpus Studies

Corpus linguistics may be viewed as a methodology, but the methodological practices adopted by corpus linguists are not uniform. This was clearly indicated in the discussion of the distinction between corpus-based and corpus-driven approaches (section 3). It was also pointed out that this distinction, superficially a methodological one, is theoretically motivated. Consequently, theoretical and methodological decisions in corpus linguistics are interlinked, although, it has to be stressed, there is no clear one-to-one correspondence between theoretical orientation and methodology in corpus studies. Therefore, the discussion of studies in this section will inevitably involve both methodological and theoretical issues.

The neo-Firthian approach to language description, and the word-based research paradigm associated with it, have indeed influenced current corpus research. However, this does not entail that all word-based studies, or studies focusing on lexical patterning, aim to contribute to a lexical description of English. Increasingly, studies investigating a grammatical phenomenon rely on the morphology or semantics of the lexis involved, while studies which focus on the collocational behavior of specific lexical items also draw on their semantic and grammatical properties. In fact, in a number of cases, the same study can be described equally well as either a grammar-focused study taking into account lexical properties, or a lexis-focused study concentrating on the grammatical behavior of specific lexical items.[11] The discussion of studies in

this section, therefore, should be read bearing in mind the indeterminacy and uncertain fusion of lexis and grammar. Corpus-based studies taking lexical items as their starting point draw on a number of theories and research approaches, and there is variation within what might, at first glance, be perceived as a single research paradigm.

Regardless of theoretical considerations, there are strong practical reasons for the appeal of word-based corpus research, even if the focus of the study is a grammatical construction. The reasons have to do with annotation. Word-based research can be carried out even with raw corpora, using software that picks out word-forms and presents the examples in a concordance (cf. Kennedy 1998: 8), although the lack of grammatical information will somehow limit the scope and effectiveness of the research. Category-based research needs, ideally, corpora annotated for grammatical structures and syntactical properties, which are time consuming to develop, or, at least, corpora annotated for the grammatical properties of words (e.g. parts of speech). However, even in a grammatically tagged corpus, it is much easier to derive concordances, say, of the verb *give* in all of its forms, than a concordance of all of the present perfect constructions. If a raw corpus is used, the former will be slightly more time consuming, but the latter will require a much bigger investment in time.[12] Halliday (1992: 64) summarizes the practical considerations of word-based and category-based research as follows:

> The lexicologist's data are relatively easy to observe: they are words, or lexical items of some kind, and . . . it is not forbiddingly difficult to parse them out. The grammarian's data are very much less accessible: I cannot even today ask the system to retrieve for me all clauses of mental process or marked circumstantial theme or high obligation modality.

A large number of corpus linguists seem to practice eclecticism in the research techniques they use, irrespective of whether they work within a specific theoretical framework, or within the research paradigm in which a given technique was first used or with which it is associated. In some respects, studies tend to adopt methodologies which demonstrate what Hunston terms a "synergy between word-based methods and category-based methods" (2002: 86).

The use of corpora in linguistic research can be placed on a cline between two points. One end treats the corpus as the sole object of study, with intuitions being excluded from all consideration as much as possible. This approach can be regarded as an extreme reaction to what Fillmore (1992) has described as "armchair linguistics," that is, the use of intuition and introspective examples as the only sources of data. The other end treats corpora as a convenient repository of instances of attested use, with the added benefit of word-and-category search and concordancing capabilities, from which the examples that fit a theory or support a point in a discussion can be selected should the user wish to do so. In the latter approach to using corpora, there is no attempt to make

the results of an experiment totally accountable to corpus data – the corpus is simply a body of casually used examples.

Between the two extreme points lie studies which combine corpus evidence with intuitions and data drawn from elicitation experiments.[13] There are perils at either end or at any point along the continuum. Those who selectively use corpora can be accused of preferring data that fits their theory while ignoring inconvenient examples. Those linguists who renounce intuition are excluding from their research a possibly rich source of evidence. Furthermore, given that the use of intuitions in the examination of the data is inescapable, a purportedly intuition-free approach to the data will involve unwitting, and therefore, unchecked, use of intuitions. Between the two extremes a blend of these criticisms may apply. Yet, from the perspective of the authors of this chapter, an approach to corpus use that combines intuition with a systematic use of corpus evidence is increasingly becoming the established norm, echoing the sentiments of Johansson (1991: 6) who cautioned that "linguists who neglect corpora do so at their peril, but so do those who limit themselves to corpora." Linguists are increasingly limiting themselves exclusively neither to corpora nor to intuition. They are using both.

In terms of their main goal, studies may be theoretically oriented, some aiming at contributing, directly or indirectly, to a specific theoretical framework. For example, studies may locate themselves within the paradigm of lexicogrammar (e.g. Hunston and Francis 2000; Renouf 2001), probabilistic grammar (e.g. Carter and McCarthy, 1999), cognitive linguistics (e.g. Gilquin 2003; Gries 2003; Gries and Stefanowitch 2004; Schmidt 2000; Schonefeld 1999), or within paradigms not readily associated with corpus-based or corpus-driven methodologies (e.g. Di Sciullo et al. 1986; Paulillo 2000). Studies may also be predominantly descriptive, that is, studies which do not explicitly subscribe to a specific theory, with an aim to discovering lexicographical and language teaching applications (e.g. McEnery and Kifle 2001, Altenberg and Granger 2002; McEnery and Xiao 2004). In terms of their research focus, studies may, for example, aim to define and explore lexical meaning (e.g. Partington, 2004), concentrate on the phraseology of a word (e.g. Hunston 2001), investigate the behavior of multi-word lexical items (e.g. De Cock et al. 1998), explore lexicogrammar (Stubbs 1996: 36), focus on the syntactic properties of grammatical structures (e.g. Duffley 2003), or examine the distribution of grammatical categories (e.g. Biber 2001). Corpus-based methodologies are also being increasingly adopted in research within pragmatics and discourse analysis (e.g. Aijmer and Stentström 2004; Archer 2005; Partington et al. 2004; Vivanco 2005; Wang 2005), critical discourse analysis (e.g. Baker 2005; Baker and McEnery 2005; Hardt-Mautner 1995; Koller and Mautner 2004; McEnery 2005; Orpin 2005; Polovina-Vukovic 2004; Sotillo and Wang-Gempp 2004), metaphor (e.g. Charteris-Black 2004; Deignan 2005), and stylistics (e.g. Burrows 2002; Semino and Short 2004; Stubbs 2005).

Within the body of corpus research it is possible to distinguish different types of studies (e.g. see Stubbs 2002: 227, 238). One way in which the studies can be categorized is in terms of their research methodology; corpus studies

can be categorized according to the extent that they rely on automatic statistical calculations or the manual examination and interpretation of concordances.[14] An example of a study that is mainly statistical would be the calculation of the frequency and strength of collocation patterns within a given span in an unannotated corpus. Although reliance on intuitions is unavoidable in the interpretation of statistical results, or the analysis of corpus examples, it can also be present in the annotation (explicit or implicit) of a corpus. A second distinction, directly related to the previous one, has to do with the size of the sample: if the study uses automatic analysis, then a large corpus can be used; if corpus examples are to be manually analyzed, then a smaller sample will have to be used.

The insights from corpus-based studies on specific areas of grammar, lexis, and their interface inform large-scale works which aim to offer a comprehensive view of the English grammar and lexicon. The next section will provide an overview of corpus-based reference grammars and dictionaries.

6 Reference Works

Corpora are now commonly used as the basis of reference grammars and dictionaries both for native speakers and learners of English. Although grammars and dictionaries are usually seen as being complementary, there has been a convergence of coverage between the two, mostly in the light of corpus evidence. Increasingly, grammars take lexical matters into account,[15] and dictionaries (usually for learners) include grammatical information in their entries. Like small-scale studies, reference works differ in the manner in, and extent to, which they make use of corpora, and the theoretical frameworks they operate in. In this section we will first discuss the impact of corpora on grammars of English before moving on to a fuller discussion of the impact of corpora on English lexicography.

Some grammars may draw their evidence and present examples from corpora only,[16] and consistently provide detailed (i.e. numerical) frequency and distributional information[17] either as part of a comprehensive grammar of English (e.g. Biber et al. 1999) or as part of a work focused on some aspect of English (e.g. the study of English verbs by Mindt 2000). Other grammars, while they take into account corpus evidence as well as findings from existing corpus-based studies, without necessarily restricting themselves to a single corpus,[18] may also draw data from elicitation experiments.[19] Such grammars provide a combination of attested and intuition-derived examples, and usually give information about frequency and distribution in more general terms (e.g. Quirk et al. 1985; Huddleston et al. 2002). Huddleston et al. (2002: 11) are quite explicit regarding their choice of data sources and their rationale for that choice:

> The evidence we use comes from several sources: our own intuitions as native speakers of the language; the reaction of other native speakers we consult when

> we are in doubt; data from computer corpora . . . and data presented in dictionaries and other scholarly work on grammar. We alternate between different sources and cross-check them against each other, since intuition can be misleading and can contain errors.

There are also a number of pedagogical grammars for students of English, either general, such as *Collins COBUILD English grammar* (1990), or with a specific focus, for example *Collins COBUILD grammar patterns 1: Verbs* (1996), and Mindt (1995), which focuses on modal verbs.

Corpus-based grammars are relatively new.[20] Dictionaries, on the other hand, have long been based upon attested language use in the form of collections of citation slips or collections of texts, for example. Some of the collections of data used to construct pre-corpus dictionaries were impressive in size given that the compilation and analysis was done manually (see Landau 2001: chs. 2 and 6). Unlike grammar-focused studies, where relatively small corpora can afford enough linguistic evidence for the purposes of the study, truly large corpora are needed for lexicographical purposes as "many words and expressions do not occur frequently enough to provide the lexicographer with enough evidence in a sample corpus" (Landau, 2001: 287). It is not a coincidence that the Bank of English (or Birmingham Corpus, as it was called originally), which was built for the purpose of compiling a dictionary (Cowie 1999; Landau 2001), is a monitor corpus, that is, an ever-expanding one. However, a representative finite corpus can also afford useful lexicographic insights. If a large representative corpus does not include a lexical item "one can conclude that the lexical item, if it exists, either is extremely uncommon or it is used almost exclusively in a specialized field that the corpus does not cover" (Landau 2001: 297). Such a corpus can also provide information about the relative frequency and distribution of lexical items and their collocation patterns, as well as grammatical information (since most representative corpora are tagged).

Lexicography has benefited from electronic corpora and its attendant software in a number of ways related to both the content and the compilation process of dictionaries. The dictionaries that utilized computer corpora very early on were English learner dictionaries, the earliest one being the *Collins COBUILD English Language Dictionary*, published in 1987. Currently, all major English learner dictionaries,[21] and, increasingly, native-speaker dictionaries[22] are corpus based (Jackson 2002: 131).

The use of corpora in dictionary construction has not merely entailed replacing citation slips with corpora. Corpora have led dictionary compilers to base decisions about inclusion and the information about entries on corpus evidence, as opposed to the more subjective decisions relied upon by the compilers of citation slip based dictionaries (cf. Landau 2001: 191–3, 205, 302–5). What Ooi (1998: 48) calls "casual citation," that is, selecting attested examples in a less than rigorous way has been supplanted by a more rigorous corpus-based approach. Modern dictionaries tend to be corpus driven rather than compiler driven. In addition, dictionaries can now provide information about frequency,

medium (written or spoken), distribution in different contexts of use, more detailed sense information, collocation patterns and grammatical properties, as a consequence of that data being available in corpora. Landau (2001: 304–5) gives an example of how a corpus can assist lexicography:

> Another perennial problem is deciding whether the present or past participle of a verb has acquired adjectival status and merits inclusion as a lemma in its own right. In the past lexicographers had no way to decide this. With a corpus that has been grammatically tagged, they do.

It must be stressed that Landau's example above should be understood as carrying the caveat that the annotation scheme, and the theoretical assumptions behind it, are known and accepted by the lexicographers involved.

Word frequencies may also be used as a criterion for including words in, or excluding words from, a dictionary. Space in a hard-copy edition of a dictionary is limited, and corpora can provide information on which to base decisions about which items to select for inclusion.[23] This is often important when compiling defining vocabularies for learner dictionaries of English. Learner dictionaries are also of interest because they are often based not only on native-speaker corpora, but also on learner corpora and on corpora comprising texts from language-teaching coursebooks. Insights from learner corpora enable dictionary compilers to provide a more detailed treatment of areas where learners seem to have problems. While corpora have contributed enormously to dictionary building, they have not replaced citation slips entirely. Corpora, particularly closed ones, cannot provide much help for dictionary makers in looking for new words. For this reason, dictionaries supplement corpus data with citations collected either manually or from the internet.

As was mentioned earlier, the availability of corpora and research tools has also provided a readily accessible testing ground for linguistic theories. In the case of dictionaries, the implications and applications of the research findings within different paradigms are helping to drive home the fact that dictionary-making is not a theory-free enterprise, and that the establishment, or influence, of different theoretical paradigms, in combination with corpus use and technological developments, will continue affecting the development of dictionaries, both in terms of the types of information dictionaries include, and the format of the dictionaries themselves. The addition of collocational and distributional information to the more traditional meanings and sense relations (usually synonymy and hyperonymy) is one of the latest developments. For example, Fillmore and Atkins (1994), working within Frame Semantics, investigated the use of the word *risk* in a corpus and compared their findings with the information given in ten monolingual dictionaries. They highlighted the following areas of difficulty: sense differentiation in the verb and noun, the distinction between 'run a risk' and 'take a risk,' and patterns of verb complementation (1994: 363), and concluded that:

> While in some cases the corpus material did suggest ways forward, more often than not it raised other more complex problems, a tough new fact to be understood and incorporated into our description of the word. It soon became apparent that the wealth of information which the corpus held could not be compressed into the format of a two-dimensional entry in a printed dictionary. The word was so complex that a multidimensional picture of it was required if we were to set out its full potential and its network of relationships within the language. (1994: 365)

Given the confrontation with data that a corpus linguist engaged in lexicographic research faces, it is hardly surprising to discover that some linguists have looked at the interface between lexis and grammar in particular, and have started to doubt that a clear division between the two exists. It appears that lexis and grammar are entangled rather than linked. For this reason, in building corpora for the construction of dictionaries, lexicographers have developed approaches to language that challenge existing linguistic categories and approaches.

Dictionary research, grammar building and lexicogrammatical research have all, clearly, been major beneficiaries of work in English corpus linguistics. Yet another related area has also benefited immensely from the development of corpora – English language teaching (ELT). This will be the subject of the next section.

7 Language Teaching

Modern approaches to the teaching of English as a foreign language have been strongly influenced by both the lexicogrammar tradition and the corpus-based approach to dictionary and grammar construction. Yet corpus use contributes to language teaching in other ways, because, apart from research on native-speaker (L1) corpora, English language teaching also benefits from research on learner corpora and corpora of ELT coursebooks (cf. Aston 2000; Aston et al. 2004; Gabrielatos 2005; Granger et al. 2002; Leech 1997b; Sinclair 2004c).

Pedagogical materials and reference books for learners can now draw on the findings of an ever-increasing and diverse body of corpus-based research. Research on native-speaker corpora has yielded a more accurate and detailed description of English, which, in turn, informs the content of pedagogical grammars and dictionaries, as well as the design of syllabuses and coursebooks (cf. Hunston and Francis 1998; Kennedy 1992; Owen 1993; Römer 2005). Research on learner corpora affords insights into the ways that learners of English use the language, provides indications about language learning processes, and contributes to second language acquisition (SLA) research (e.g. Granger et al. 2002; Jones and Murphy 2005). The identification of frequent learner problems, particularly problems specific to learners of a given first language, can further facilitate the design of syllabuses and pedagogical materials (e.g. Nesselhauf 2005). Corpora of English language teaching coursebooks can provide a helpful

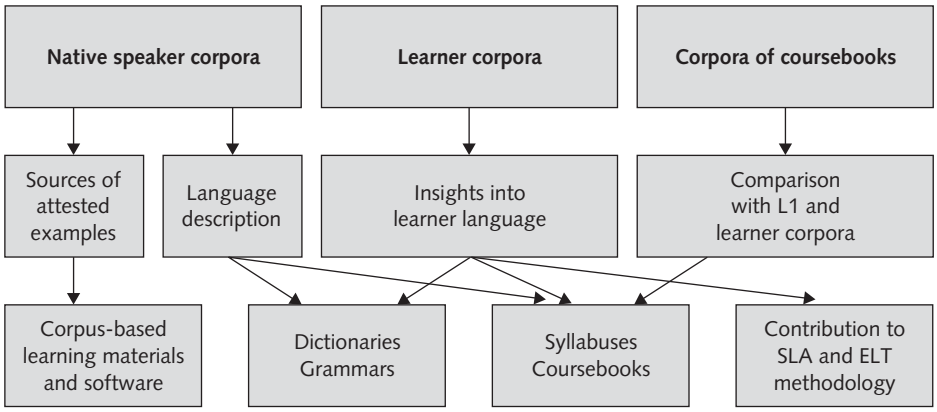

Figure 3.1 Corpora and language teaching

comparison between native language use in different contexts and the language that learners are exposed to in coursebooks (e.g Harwood 2005; Römer 2004). The analysis of discrepancies will provide a helpful guide as to the kinds of texts that should be included in pedagogical materials. Such corpora already guide decisions about the content and focus of dictionaries (e.g. the *Macmillan English Dictionary for Advanced Learners*, 2002).[24] The examination of corpora of coursebooks can also reveal whether, and to what extent, the language in coursebook texts influences the speech and writing of learners. Finally, language corpora can be used by learners and teachers as a source of attested language examples, with learners either having access to corpora and having been familiarized with corpus software, or working with printouts of concordances (e.g. Johns 1991; Aston 1997). Figure 3.1 (adapted from Gabrielatos 2003) provides an outline of the contribution of corpora and corpus research to language teaching, as summarized above, and shows the multiple connections between different types of corpora, and the insights that their analysis affords, and applications to language learning and teaching. The remainder of this section expands on the contribution of corpus linguistics to language teaching and learning.

Research on L1 corpora has yielded convincing evidence that traditional, intuition-based views on language are very often at odds with actual language use (e.g. Sinclair 1997: 32–4). Also, corpus-based research on general and specialized corpora has revealed patterns and uses that introspective accounts had previously failed to detect. This is pertinent to language teaching, as the information about English structure and use that is communicated to learners, either by pedagogical materials or teachers, is still, to a large extent, based on intuitions. As we have already mentioned, intuitions are useful, but not necessarily accurate. Having a native or good command of a language does not endow a language teacher with a conscious, clear and comprehensive picture of the language in all of its contexts of use. What is more, native

speaker intuitions vary from user to user. A case in point is the published view of a native-speaker teacher that in English, "question tags, along with bowler hats, mostly belong to 1960s BBC broadcasts" (Bradford 2002: 13). This view is in sharp contrast with the corpus findings of Biber et al. (1999: 211), who report that "about every fourth question in conversation is a question tag." However, as was shown in section 6 above, a large number of pedagogical reference books which are informed by corpus studies are now available to English language learners.

Studies of learner language mainly compare learner use of specific features in different contexts with that of native-speakers in quantitative and qualitative terms, and engage with the examination and classification of learner errors. Error analysis seeks to identify frequent errors or error patterns with reference to one or more of the following factors: the learners' L1, level and age, the medium of production (speech or writing), the genre, and the context of use. Studies of learner language, usually based on written corpora, have also focused on learner use of a large variety of features of lexicogrammar, such as lexical chunks (De Cock et al. 1998), complement clauses (Biber and Reppen 1998), the progressive aspect and questions (Virtanen 1997, 1998), and the use of epistemic modality (McEnery and Kifle 2002), as well as discourse features, such as overstatement (Lorenz 1998), connectors (Altenberg and Tapper 1998), and speech-like elements in writing (Granger and Rayson 1998). Corpus-based research of learner language contributes to English language teaching in two respects. Research findings point towards the aspects of learner use which should be prioritized in language instruction and aid the compilation of pedagogical and reference materials at different levels of competence. The examination of learner language also affords insights into the process of language learning (e.g. Tono 2000).

Both native-speaker and learner corpora can be used directly in language teaching, either in class or for self-study. The former can provide exposure to language in use, whereas the latter can raise awareness of language problems common to a specific L1. Corpus-based approaches to language awareness can be distinguished according to two ways that a corpus is utilized (Leech 1997b: 10). In the first, the corpus is used as a source of attested language examples for the teacher or materials writer (e.g. Tribble and Jones 1990; Tribble 1997; Granger and Tribble 1998; Osbourne 2000). Johns (2002: 108–9) provides an example of the use of concordances in the classroom: learners are given ten groups of five concordance lines each where the key word (in this case a noun) is missing, as well as a list of the ten missing nouns (each corresponding to one group of concordance lines), and are asked to decide which noun completes each group. Teachers can manipulate the corpus examples by restricting them to a specific medium (writing/speech), genre, or text type. Of course, when using very small or selective corpus samples, teachers need to inform learners that no valid conclusions can be drawn about the actual or relative frequency of a language feature on the basis of the corpus examples. Teachers

can also regulate the amount of text available to learners, from only a few words on either side of the key word to a sentence or a paragraph. In the second approach, learners work directly with corpora (cf. Aston 1996), either following instructions given by the teacher or contained within a CALL[25] program (e.g. Hughes 1997; Milton 1998), or working on areas of their own choice (e.g. Johns 1997). Bernardini (2002: 174–5) reports on an example of learners influencing the direction of a corpus-based lesson. The learners were investigating the phraseology of *high standards* in the BNC, using the concordance function to look for typical collocates. While examining verb collocates on the left-hand side, a learner noticed the intensifier *extremely*. This led the learner to query the phrase *extremely high* and investigate right-hand collocating nouns. The teacher used this opportunity to ask learners to investigate the distribution of *extremely high* in the subcorpora.

Corpus use in English language teaching is often associated with a "data-driven" approach to learning, which regards the learner as a researcher (Johns 1991). However, it should be pointed out that corpus use is not restricted to any single teaching methodology. It is compatible with all methodologies that accept an explicit focus on language structure and use, i.e. teaching approaches which see a role, central or marginal, for consciousness-raising through noticing (e.g. Sharwood Smith 1981; Lightbown 1985; Schmidt 1990), that is, encouraging and guiding learners to pay attention to language features and patterns and, not unlike corpus researchers, formulate and test generalizations themselves, rather than being given a rule. In other words, corpus use fits equally well within language-based and task-based approaches to language learning (cf. Nunan 1989; Fotos and Ellis 1991; Loschky and Bley-Vroman 1993; Skehan 1998).

The use of corpora in language teaching has provided new opportunities for learner independence. According to Johns (1997: 101), when using corpora or corpus-based materials, "students define their own tasks as they start noticing features of the data for themselves – at times features that had not previously been noticed by the teacher." Also, corpus use has given a new lease of life to the language lab, and has suggested a more flexible and learner-centered use for CALL materials (e.g. McEnery et al. 1997). The introduction of corpora to the language classroom has also challenged the traditional role of teachers, which does not mean that the teacher's role is diminished; rather, that it is enriched and diversified. The teacher is seen not so much as the provider of facts about language as a consultant or co-researcher.

Another benefit of working with corpus samples from representative corpora of different varieties (e.g. British or American English) and different genres (e.g. academic English, chatroom English) is that learners develop an awareness of different varieties of English in a number of contexts. This exposure is expected to facilitate their understanding and enrich their language use, but, more importantly, to drive home the fact that English is anything but uniform.

8 Language Change

So far, the discussion of the contribution of corpora to linguistic research has been concerned with issues relating to the description and analysis of modern English. However, the availability, and relative ease of construction, of corpora comprising texts from different periods of the development of English[26] makes it possible to investigate changes in different aspects of the English language. Renouf (1997: 185) points out that "historical text study has long been ripe for automation." Electronic corpora have made data collection more time-efficient and have enabled researchers to tackle areas hitherto made forbidding by the volume of materials that needed to be collected (Rissanen 1997: 6). Nevalainen and Raumolin-Brunberg (2003: 23) mention a further advantage of corpus-based diachronic studies, namely that the language is produced without the prompting, participation or presence of a researcher, which "may affect the linguistic choices people make."

Studies on the development of English usually examine four periods: Old, Middle, Early Modern and Modern English. In terms of what is compared, we can distinguish between (1) contrasting different historical periods; (2) contrasting one or more past periods and Modern English; and (3) given that computer corpora on Modern English now span more than forty years, tracking changes within recent decades. What follows is a brief review of studies providing a range of examples of the use of corpora in the study of long-term and recent language change.

Kytö (1996) compared the morphology of adjective comparison (inflectional, periphrastic, and double forms)[27] in Late Middle and Early Modern English using the Helsinki Corpus of English Texts. The results show a slight increase in the inflected forms, and a slight decrease in the periphrastic forms, with the changes being more pronounced in the case of the superlative, and only sporadic use of the double form. Lopez-Couso and Mendez-Naya (2001) examined the development of declarative complement clauses with *if* and *though* in Old, Middle and Early Modern English, using the diachronic part of the Helsinki Corpus of English Texts. According to their data, the frequency of *if* complements has increased since the Old English period, whereas *though* complements, although more frequent in Old English, became obsolete in the early seventeenth century.

Krug (2000) studied the development of the modal expressions *have got to/gotta*, *have to/hafta* and *want to/wanna* in a number of diachronic and contemporary corpora.[28] Apart from concluding that they all show signs of ongoing auxiliarization, that is, they now display more of the formal characteristics of modal auxiliaries, he also observed that "frequency seems to be a fundamental parameter in the genesis of the new category [i.e. the modal expressions *have got to/gotta*, *have to/hafta* and *want to/wanna*]" (2000: 251). Mair et al. (2003) compared the tag frequencies in two corpora, LOB (1961) and F-LOB (1991), to investigate whether English has become more 'nominal.' They found that nouns,

particularly proper nouns, and adjectives were significantly more frequent in FLOB, as were 'noun + common noun' sequences. Leech (2003) and Smith (2003) examined a number of British and American English corpora from 1961 and 1991/2[29] and showed that there was a decline overall in the use of central modals, and an increase in the frequency of semi-modals.[30]

A new corpus developed by Bas Aarts and Sean Wallis at University College London is the *Diachronic Corpus of Present-day Spoken English* (DCPSE).[31] DCPSE contains spontaneous spoken British English, and comprises comparable categories from the *London-Lund Corpus* (1960–76) and the British English component of the *International Corpus of English* (ICE-GB, early 1990s). The DCPSE contains 800,000 words, and has been grammatically annotated (tagged and parsed), as well as annotated for features of interaction, such as speaker and turn overlaps, making it particularly suitable for diachronic research on spoken grammar.

Baayen and Renouf (1996) focused on neologisms and compared the productivity of two prefixes (*un-*, *in-*) and three affixes (*-ly*, *-ness*, *-ity*) in *The Times* database (1989–92, 80 mil. words) and in the COBUILD corpus (18 mil. words). One of their conclusions was that "word formation in the native stratum of the lexicon is much more productive than dictionaries would suggest" (1996: 92). Collier (1998) outlines a two-stage methodology for tracking changes in semantic relations, on the evidence of collocational profiles, in the section of UK newspapers in the ACRONYM corpus database system (Renouf 1996). Two databases of significant collocations are compared: the year in focus and the previous years. Collocational changes are tagged according to whether their significance has increased/decreased (termed *up/down* collocates), or whether they have appeared/disappeared (*new/gone* collocates). In the second stage, the significant collocates of the target word are extracted, but only those fulfilling certain criteria, say the *up* and *new* collocates, are considered. The *up* and *new* collocates are then treated as nodes and their collocates are calculated, but only those occurring more than once are retained. The common collocates-of-collocates are used to draw the semantic profile of the node, as their "collocate profiles overlap to a lesser or greater extent with that of the original target word" (Collier 1998: 264). It is interesting to note that using newspaper corpora enables the researchers to identify extremely short-term changes, albeit in a specific domain.

One facet of language change which lies on the lexicogrammatical interface is grammaticalization.[32] Grammaticalization can be seen diachronically as "a process whereby lexical items and constructions come in certain linguistic contexts to serve grammatical functions, and, once grammaticalized, continue to develop new grammatical functions" (Hopper and Traugott 1993: xv).

Rissanen (1997) traced the emergence and development of the pronominalization of *one* from Old to Modern English, and examined six types of pronominal uses, based mainly on the Helsinki Corpus.[33] According to his analysis, developments were more pronounced in Middle and Early Modern English. Rissanen (1997: 135) sees a connection between the pronominalization of *one*

and the "loss of the inflectional endings of the language and the consequent collapse of the Old English case system." A study by Brems (2003), focussing on the syntactical properties of 'measure nouns' in *of*-phrases (e.g. 'a kilo of apples'), investigates one of the parameters defining grammaticalization, *coalescence*, "a syntactic criterion [which] concerns an increase in bondedness or syntactic cohesion of the elements that are in the process of grammaticalizing, i.e. what were formerly individually autonomous signs become more dependent on each other to the extent that they are increasingly interpreted as together constituting one "chunk," which as a whole expresses a (grammatical) meaning" (2003: 291). The study revealed that although some constructions, namely *bunch(es) of*, *heap(s) of* and *piles(s) of*, "have developed a quantifier use comparable to that of regular quantifiers," not all measure noun constructions show the same degree of grammaticalization (p. 309). A theoretically important observation was that the assessment of the structural status of measure nouns in these constructions was made difficult by the interdependence of their lexical and grammatical status (2003).

The examples of studies on language change presented here testify to the wealth of opportunities afforded by corpora to examine a very wide span of the history of English. Furthermore, the variety of theoretical approaches taken to the diachronic study of English is a further indication that, while the lexical approach is an important and major contribution to English corpus linguistics, it represents but one way to approach corpus data.

Conclusion

In this chapter we have not tried to present a comprehensive overview of English corpus linguistics, as the scope of corpus based studies of English is vast. What we have done instead is to outline the major impacts that corpora have had on the study of the English language. From changing the way in which basic reference resources relating to the English language have been developed through to the development of a critical, lexicogrammatical approach to establish theories and categorization of language, and beyond lexicogrammar to the study of English through the ages using a range of theoretically informed approaches, corpus data has changed the way the English language is studied. It has also changed the way that the language is taught. So while the term 'English corpus linguistics' will remain somewhat vague and inclusive, covering potentially any study of the English language which uses corpus data, this chapter has presented those changes to the study of English which it would be difficult to imagine occurring had English language corpora not been developed.

Appendix 1 Information on corpora mentioned in the chapter

ARCHER Corpus (A Representative Corpus of Historical English Registers)

Language variety	*British and American English*
Size	1.7 million words
Medium	Writing, including written representation of speech (e.g. drama)
Time period	Early Modern English (1650–1990)
Annotation	Unannotated
More information	Biber et al. (1994a, 1994b)

The Bank of English

Language variety	*British English*
Size	450 million words in January 2002
Medium	Writing
Time period	Mostly after 1990
Annotation	POS tagged
More information	User guide: www.titania.bham.ac.uk/docs/svenguide

The British National Corpus (BNC)

Language variety	*British English*
Size	100 million words
Medium	90% writing, 10% speech
Time period	Early 1990s
Annotation	Part of speech tagging using CLAWS 5
More information	Aston and Burnard (1998), BNC website: www.natcorp.ox.ac.uk

The Brown University Corpus (Brown)

Language variety	*American English*
Size	1 million words
Medium	Writing
Time period	1960
Annotation	Raw version and different annotated versions
More information	Francis and Kucera (1964), Manual: http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM

The Freiburg–Brown Corpus of American English (Frown)

Language variety	*American English*
Size	1 million words
Medium	Writing
Time period	1991
Annotation	
More information	Hundt et al. (1999), http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM

The Freiburg–LOB Corpus of British English (FLOB)

Language variety	*British English*
Size	1 million words
Medium	Writing
Time period	1991
Annotation	POS tagged using CLAWS 8
More information	Hundt et al. (1998), http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM

The Helsinki Corpus (Diachronic Part)

Language variety	*British English*
Size	1.5 million words
Medium	Writing
Time period	Old, Middle and Early Modern English (c.750 to c.1700)
Annotation	Unannotated
More information	Kytö (1996)

The International Corpus of English, British English component (ICE–GB)

Language variety	*British English*
Size	1 million words
Medium	Writing and speech
Time period	1990–1998
Annotation	POS tagged and parsed
More information	Greenbaum (1996); www.ucl.ac.uk/english-usage/ice-gb/index

The Lancaster/Oslo–Bergen Corpus (LOB)

Language variety	*British English*
Size	1 million words
Medium	Writing
Time period	1961
Annotation	Raw version and POS tagged version using CLAWS 1
More information	Johansson et al. (1978) and Johansson et al. (1986)

The Survey of English Usage

Language variety	*British English*
Size	1 million words
Medium	Writing and speech
Time period	Between 1955 and 1985
Annotation	POS tagged
More information	www.ucl.ac.uk/english-usage/about/history

NOTES

1 See also Meyer and Nelson, ch. 5, this volume, for a discussion of data collection and corpus building.

2 Annotation is the manual or automatic process of adding information to a corpus. The information may refer to the grammatical, syntactical, semantic, or pragmatic properties of words, phrases, structures, sentences, or longer stretches of text. Grammatical annotation is also referred to as (grammatical or part-of-speech) *tagging*.

3 Perhaps 'to insulate the theory from the data' describes this practice more clearly.

4 See section 4, 'Lexicogrammar and lexical grammar,' for a more detailed treatment of this approach.

5 For example, Biber et al. (1999) make use of some frameworks used in Quirk et al. (1985), but they are also influenced by research in lexicogrammar (Biber et al. 1999: viii, 13).

6 For a discussion of statistical collocational analysis see Barnbrook (1996: ch. 5), Hunston (2002: ch. 4).

7 See McEnery (2003) for a further discussion of these criticisms of corpus annotation.

8 For example, Phillips (1989) lemmatizes on the basis of a preliminary investigation of collocation patterns.

9 The number of words that the term 'large corpus' denotes has been constantly increasing. The one-million-word Brown Corpus was considered large in the mid-1960s, whereas, forty years later, the Bank of English is almost half a billion words.

10 See section 5 for a brief discussion of the practical appeal of word-based research.

11 For example, Kennedy (1998: 121–54) presents within "grammatical studies centred on morphemes or words" (1998: 121) research focusing on the frequency of modal verbs, verb+particle combinations, prepositions, and conjunctions, together with research on tense-aspect marking, voice, and the subjunctive.

12 Note, however, that research into querying grammatically parsed corpora is developing apace; see, for example, Nelson et al. (2002).

13 For a discussion of data collection see Meyer and Nelson, ch. 5, this volume.

14 See Hunston (2002: chs. 3 and 4) for a discussion.
15 Biber et al. (1999: 13–14) make it explicit that they treat grammatical and lexico-grammatical patterns.
16 Biber et al. (1999) is based on a single corpus, the Longman Spoken and Written English Corpus (40 million words); Mindt (2000) is based on the British National Corpus (BNC) (Aston and Burnard 1998).
17 It is, of course, feasible to provide frequency and distributional information even when the book is based on studies carried out using different corpora, particularly when the corpora represent specialized domains. If different general corpora are used, this will assume that the corpora are comparable in terms of representativeness and size.
18 Huddleston and Pullum (2002) use the Brown Corpus, the Australian Corpus of English, the LOB corpus and the Wall Street Journal Corpus, as well as data from newspapers, plays, books, and film scripts (2002: 11, n. 3); Quirk et al. (1985) is informed by research using the Survey of English Usage, the Brown Corpus, and the LOB corpus (1985: 33).
19 For a discussion, see Meyer and Nelson, ch. 5, this volume.
20 Though there are grammars dating back some time which are clearly corpus based, most notably the grammar of Fries (1952).
21 *Cambridge Advanced Learner's Dictionary* (2003, 2nd edn.), *Collins COBUILD Advanced Learner's English Dictionary* (2003, 4th edn.), *Longman Dictionary of Contemporary English* (2003, 4th edn.), *Macmillan English Dictionary for Advanced Learners* (2002), *Oxford Advanced Learner's Dictionary* (2002, 6th edn.).
22 *Collins English Dictionary* (2003, 6th edn.), *Oxford Dictionary of English* (2003, 2nd edn.).
23 However, space in the CD-rom editions of dictionaries is much less restricted, and in online dictionaries space is almost unlimited.
24 The Macmillan Curriculum Corpus is "a 20 million-word corpus specially developed for the *Macmillan School Dictionary*. This unique corpus includes texts from coursebooks of different levels and school subjects, from countries where English is used as a second language, and from countries where English is the medium of instruction in schools" (www.macmillandictionary.com/school/about/corpus).
25 Computer Assisted Language Learning.
26 Notably the diachronic component of the Helsinki Corpus (Kytö and Rissanen 1988).
27 Kytö (1988: 124) provides the example of *easy*, which, during the Middle and Early Modern English periods, appeared in all three forms: inflectional (*easier/easiest*), periphrastic (*more/most easy*), and double (*more easier/most easiest*).
28 The diachronic component of the Helsinki Corpus (Old to Early Modern English, pre-850 to ca. 1700), the ARCHER Corpus (1650–1990), as well as a corpus of Shakespeare's works, LOB/FLOB, Brown/Frown, BNC, and the *Guardian* CD-rom (1990–7).
29 Brown (1961), LOB (1961), FLOB (1991), Frown (1992), Survey of English Usage (1959–85), ICE-GB (1990–2).
30 See also Biber (2004); Mair and Leech (ch. 14, this volume).
31 www.ucl.ac.uk/english-usage/diachronic/index.
32 See Lindquist and Mair (2004).
33 Rissanen (1997: 88) notes that "other corpora, concordances, dictionaries and primary texts have also been studied."

FURTHER READING

Introductory books

Barnbrook, G. (1996) *Language and computers: a practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Biber, D., Conrad, S., and Reppen, R. (1998) *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Hunston, S. (2002) *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Kennedy, G. (1998) *An introduction to corpus linguistics*. London: Longman.

McEnery, T. and Wilson, A. (2001) *Corpus linguistics*, 2nd edn. Edinburgh: Edinburgh University Press.

Meyer, C. (2002) *English corpus linguistics: an introduction.* Cambridge: Cambridge University Press.

Partington, A. (1996) *Using corpora for English language research and teaching*. Amsterdam: John Benjamins.

Sampson, G. (2001) *Empirical linguistics*. London: Continuum.

Stubbs, M. (1996) *Text and corpus analysis: computer assisted studies of language and culture*. Oxford: Blackwell.

Stubbs, M. (2001) *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

Tognini-Bonelli, E. (2001) *Corpus linguistics at work.* Amsterdam and Philadelphia: John Benjamins.

Edited volumes

Aijmer, K. and Altenberg, B. (eds.) (2004) *Advances in corpus linguistics.* Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23), Göteborg, May 22–6, 2002. Amsterdam: Rodopi.

Altenberg, B. and Granger, S. (eds.) (2002) *Lexis in contrast: corpus-based approaches*. Amsterdam: John Benjamins.

Botley, S., McEnery, T., and Wilson, A. (eds.) (2000) *Multilingual corpora in teaching and research.* Amsterdam: Rodopi.

Burnard, L. and McEnery, T. (eds.) (2000) *Rethinking language pedagogy from a corpus perspective*. Frankfurt am Main: Peter Lang.

Connor, U. and Upton, T. A. (eds.) (2004) *Applied corpus linguistics: a multidimensional perspective.* Amsterdam: Rodopi.

Garside, R., Leech, G., and McEnery, T. (eds.) (1997) *Corpus annotation: linguistic information from computer text corpora*. London; New York: Longman.

Granger, S., Hung, J., and Petch-Tyson, S. (eds.) (2002) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Granger, S. and Petch-Tyson, S. (eds.) (2003) *Extending the scope of corpus-based research: new applications, new challenges*. Amsterdam: Rodopi.

Ketteman, B. and Marko, G. (eds.) (2002) *Teaching and learning by doing corpus analysis*. Proceedings from the Fourth International Conference on Teaching and Language Corpora, Graz, July 19–24, 2000. Amsterdam: Rodopi.

Lindquist, H. and Mair, C. (2004) *Corpus approaches to grammaticalization in English*. Amsterdam: John Benjamins.

Leistyna, P. and Meyer, C. F. (eds.) (2003) *Corpus analysis: language structure and language use*. Amsterdam: Rodopi.

Mair, C. and Hundt, M. (eds.) (2000) *Corpus linguistics and linguistic theory* (ICAME 20) Amsterdam: Rodopi.

Renouf, A. (ed.) (1998) *Explorations in corpus linguistics*. Amsterdam: Rodopi.
Scott, M. and Thompson, G. (eds.) (2001) *Patterns of text: in honour of Michael Hoey*. Amsterdam and Philadelphia: John Benjamins.
Simpson, R. C. and Swales, J. M. (eds.) (2001) *Corpus linguistics in North America*. Ann Arbor: University of Michigan Press.

Lexical approach

Hoey, M. (1991) *Patterns of lexis in text*. Oxford: Oxford University Press.
Hunston, S. and Francis, G. (1999) *Pattern grammar*. Amsterdam: John Benjamins.
Sinclair, J. McH. (1991) *Corpus concordance collocation*. Oxford: Oxford University Press.

REFERENCES

Aarts, J. (2002) Review of E. Tognini-Bonelli, *Corpus linguistics at work*. *International Journal of Corpus Linguistics* 7 (1), 118–23.
Aijmer, K. and Altenberg, B. (eds.) (1991) *English corpus linguistics: studies in honour of Jan Svartvik*. London: Longman.
Aijmer, K. and Altenberg, B. (eds.) (2004) *Advances in corpus linguistics*. Papers from the 23rd International Conference on English Language Research on Computerized Corpora (ICAME 23), Göteborg, May 22–6, 2002. Amsterdam: Rodopi.
Aijmer, K. and Stenström, A.-B. (eds.) (2004) *Discourse patterns in spoken and written corpora*. Amsterdam: John Benjamins.
Alatis, J. (ed.) (1991) *Georgetown University round table on languages and linguistics 1991*. Washington, DC: Georgetown University Press.
Altenberg, B. and Granger, S. (eds.) (2002) *Lexis in contrast: corpus-based approaches*. Amsterdam: John Benjamins.
Altenberg, B. and Tapper, M. (1998) The use of adverbial connectors in advanced Swedish learners' written English. In Granger (ed.), 80–93.
Archer, D. (2005) *Questions and answers in the English courtroom (1640–1760: a sociopragmatic analysis*. Amsterdam: John Benjamins.
Aston, G. (1996) The British National Corpus as a language learner resource. Paper presented at the Second Conference on Teaching and Language Corpora, Lancaster University, UK, August 9–12. Also online: www.natcorp.ox.ac.uk/using/papers/aston96a.
Aston, G. (1997) Enriching the learning environment: corpora in ELT. In Wichmann et al. (eds.), 51–64.
Aston, G. (2000) Corpora and language teaching. In Burnard and McEnery (eds.), 7–17.
Aston, G. and Burnard, L. (1998) *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.
Aston, G., Bernardini, S., and Stewart, D. (eds.) (2004) *Corpora and language barriers*. Amsterdam and Philadelphia: John Benjamins.
Atkins, B. T. S. and Zampolli, A. (eds.) (1994) *Computational approaches to the lexicon*. Oxford: Oxford University Press.
Baayen, R. H. and Renouf, A. (1996) Chronicling the times: productive lexical innovations in an English newspaper. *Language: Journal of the*

Linguistic Society of America 72 (1), 69–96.

Baker, M., Francis, G., and Tognini-Bonelli, E. (eds.) (1993) *Text and technology: in honour of John Sinclair*. Philadelphia and Amsterdam: John Benjamins.

Baker, P. (2005) *Public discourses of gay men*. London: Routledge.

Baker, P. and McEnery, T. (2005) A corpus-based approach to discourses of refugees and asylum seekers in UN and newspaper texts. *Language and Politics* 4 (2), 197–226.

Barnbrook, G. (1996) *Language and computers: a practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Bazell, C. E., Catford, J. C., Halliday, M. A. K., and Robins, R. H. (eds.) (1966) *In memory of F. R. Firth*. London: Longman.

Bernandini, S. (2002) Exploring new direction for discovery learning. In Ketteman and Marko (eds.), 165–82.

Biber, D. (2001) Using corpus-based methods to investigate grammar and use: some case studies on the use of verbs in English. In Simpson and Swales (eds.), 101–15.

Biber, D. (2004) Modal use across registers and time: an analysis based on the ARCHER corpus. In A. A. Curzan and K. Emmons (eds.), *Studies in the history of the English language II: unfolding conversations*. Berlin: Mouton de Gruyter, 189–216.

Biber, D., Finegan, E., and Atkinson, D. (1994a) ARCHER and its challenges: Compiling and exploring a representative corpus of historical English registers. In Fries et al. (eds.), 1–13.

Biber, D., Finegan, E., Atkinson, D., Beck, A., Burges, D., and Burges, J. (1994b) The design and analysis of the ARCHER Corpus: a progress report. In Kytö et al. (eds.), 3–6.

Biber, D., Conrad, S., and Reppen, R. (1998) *Corpus linguistics: investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan E. (1999) *Longman grammar of spoken and written English.* London: Longman.

Biber, D. and Reppen, R. (1998) Comparing native and learner perspectives on English grammar: a study of complement clauses. In Granger (ed.), 145–58.

Boswood, T. (ed.) (1997) *New ways of using computers in language teaching*. Alexandria, VA: TESOL.

Bradford, R. (2002) Grammar is by statisticians, language is by humans. *IATEFL Issues* 167, 13.

Brems, L. (2003) Measure noun constructions: an instance of semantically-driven grammaticalization. *International Journal of Corpus Linguistics* 8 (2), 283–312.

Burnard, L. and McEnery, T. (eds.) (2000) *Rethinking language pedagogy from a corpus perspective*. Papers from the third international conference on teaching and language corpora. Hamburg: Peter Lang.

Burrows, J. (2002) The Englishing of Juvenal: computational stylistics and translated texts. *Style* 36 (4), 677–9.

Cambridge advanced learner's dictionary (2003) 2nd edn. Cambridge: Cambridge University Press.

Carter, R. and McCarthy, M. (1999) The English get-passive in spoken discourse: description and implications for an interpersonal grammar. *English Language and Linguistics* 3 (1), 41–58.

Charteris-Black, J. (2004) *Corpus approaches to critical metaphor analysis*. Basingstoke: Palgrave-Macmillan.

Collier, A. (1998) Identifying diachronic change in semantic relations. In Renouf (ed.), 259–68.

Collins COBUILD Advanced learner's English dictionary (2003) 4th edn. London: Harper Collins.
Collins COBUILD English grammar (1990) London: Harper Collins.
Collins English dictionary (2003) 6th edn. London: Harper Collins.
Collins COBUILD Grammar patterns 1: verbs (1996) London: Harper Collins.
Cowie, A. P. (1999) *English dictionaries for foreign learners: a history*. Oxford: Oxford University Press.
Crookes, G. and Gass, S. M. (eds.) (1993) *Tasks and language learning: Integrating theory and practice*. Clevedon: Multilingual Matters.
De Cock, S., Granger, S., Leech, G., and McEnery, A. M. (1998) An automated approach to the phrasicon of EFL learners. In Granger (ed.), 67–79.
Deignan, A. (2005) Metaphor and corpus linguistics. *Converging Evidence in Language and Communication Research 6*. Amsterdam: John Benjamins.
Di Sciullo, A.-M.; Muysken, P., and Singh, R. (1986) Government and code-mixing. *Journal of Linguistics* 22 (1), 1–24.
Duffley, P. J. (2003) The gerund and the *to*-infinitive as subject. *Journal of English Linguistics* 31 (4), 324–52.
Facchineti, R. and Krug, M. (eds.) (2003) *Modality in contemporary English*. Berlin: Mouton de Gruyter.
Fillmore, C. J. (1992) "Corpus linguistics" or "Computer-aided armchair linguistics." In J. Svartvik (ed.), 35–60.
Fillmore, C. J. and Atkins, B. T. S. (1994) Starting where the dictionaries stop: The challenge of corpus lexicography. In Atkins and Zampolli (eds.), 349–93.
Firth, J. R. (1951/1957) Modes of meaning. In *Papers in Linguistics 1934–1951*. London: Oxford University Press.
Firth, J. R. (1968) A synopsis of linguistic theory. In Palmer (ed.), 168–205.
Flowerdew, J. (ed.) (2001) *Academic discourse.* London: Longman.
Fotos, S. and Ellis, R. (1991) Communicating about grammar: a task-based approach. *TESOL Quarterly* 25 (4), 605–28.
Francis, W. N. and Kučera, H. (1964) *A standard corpus of present-day edited American English.* Providence: Brown University.
Fries, C. (1952) *The structure of English*. New York: Harcourt Brace.
Fries, U., Tottie, G., and Schneider, P. (eds.) (1994) Creating and using English language corpora. Papers from the Fourteenth International Conference on English Language Research on Computerized Corpora, Zürich 1993. Amsterdam: Rodopi.
Gabrielatos, C. (2003) Corpora and ELT: Just a fling, or the real thing? Plenary address at INGED 2003 International Conference, Multiculturalism in ELT Practices: Unity and Diversity, organized jointly by BETA (Romania), ETAI (Israel), INGED (Turkey), and TESOL Greece, Baskent University, Ankara, Turkey, October 10–12, 2003.
Gabrielatos, C. (2005) Corpora and language teaching: just a fling or wedding bells? *TESL-EJ* 8 (4) www.tesl-ej.org/ej32/al.
Garside, R., Leech, G., and McEnery, T. (eds.) (1997) *Corpus annotation: linguistic information from computer text corpora*. London: Longman.
Gilquin, G. (2003) Causative *get* and *have*: so close, so different. *Journal of English Linguistics* 31 (2), 125–48.
Granger, S. (ed.) (1998) *Learner English on computer*. London: Addison Wesley Longman.
Granger, S. and Rayson, P. (1998) Automatic profiling of learner texts. In Granger (ed.), 119–31.
Granger, S. and Tribble, C. (1998) Learner corpus data in the foreign language classroom: form-focused instruction and data-driven learning. In Granger (ed.), 199–209.

Granger, S., Hung, J., and Petch-Tyson, S. (eds.) (2002) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: John Benjamins.

Greenbaum, S. (ed.) (1996) *Comparing English worldwide: the International Corpus of English*. Oxford: Clarendon Press.

Gries, S. T. (2003) Towards a corpus based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1, 1–27.

Gries, S. T. and Stefanowitsch, A. (2004) Extending collostructional analysis: a corpus-based perspective on 'alternations.' *International Journal of Corpus Linguistics* 9 (1), 97–129.

Halliday, M. A. K. (1966) Lexis as a linguistic level. In Bazell et al. (eds.), 148–62.

Halliday, M. A. K. (1991) Corpus studies and probabilistic grammar. In Aijmer and Altenberg (eds.), 30–40.

Halliday, M. A. K. (1992) Language as system and language as instance: the corpus as a theoretical construct. In Svartvik (ed.), 61–77.

Hardt-Mautner, G. (1995) Only connect: critical discourse analysis and corpus linguistics. UCREL Technical Papers 6. Lancaster University.

Harwood, N. (2005) What do we want EAP teaching materials for? *Journal of English for Academic Purposes* 4 (2), 149–61.

Hoey, M. (1991) *Patterns of lexis in text*. Oxford: Oxford University Press.

Hoey, M. (1997) From concordance to text structure: new uses for computer corpora. In Lewandowska-Tomaszczyk and Melia (eds.), 2–23.

Hopper, P. J. and Traugott, E. C. (1993) *Grammaticalization.* Cambridge: Cambridge University Press.

Huddleston, R. and Pullum, G. K., et al. (2002) *The Cambridge grammar of the English language*. Cambridge: Cambridge University Press.

Hughes, G. (1997) Developing a computing infrastructure for corpus-based teaching. In Wichmann et al. (eds.), 292–307.

Hundt, M., Sand, A., and Siemund, R. (1998) *Manual of information to accompany the Freiburg-LOB Corpus of British English ('FLOB')* Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg.

Hundt, M., Sand, A., and Skandera, P. (1999) *Manual of information to accompany the Freiburg-Brown Corpus of American English ('Frown')* Freiburg: Englisches Seminar, Albert-Ludwigs-Universität Freiburg.

Hunston, S. (2001) Colligation, lexis, pattern and text. In Scott and Thompson (eds.), 13–33.

Hunston, S. (2002) *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. and Francis, G. (1998) Verbs observed: a corpus-driven pedagogic grammar. *Applied Linguistics* 19 (1), 45–72.

Hunston, S. and Francis, G. (2000) *Pattern grammar*. Amsterdam: John Benjamins.

Hyltenstam, K. and Pienemann, M. (eds.) (1985) *Modelling and assessing second language acquisition*. Clevedon, North Somerset: Multilingual Matters.

Jackson, H. (2002) *Lexicography: an introduction*. London: Routledge.

Johansson, S. (1991) Computer corpora in English language research. In Johansson and Stenström (eds.), 3–6.

Johansson, S., Leech, G., and Goodluck, H. (1978) *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*. Department of English, University of Oslo.

Johansson, S., Atwell, E., Garside, R., and Leech, G. (1986) *The tagged LOB Corpus: user's manual*. Norwegian Computing Centre for the Humanities, Bergen.

Johansson, S. and Stenström, A.-B. (eds.) (1991) *English computer corpora: selected papers and research guide*. Berlin: Mouton de Gruyter.

Johns, T. (1991) Should you be persuaded: two examples of data driven learning. In Johns and King (eds.), 1–16.

Johns, T. (1997) Contexts: the background, development and trialling of a concordance-based CALL program. In Wichmann et al. (eds.), 100–15.

Johns, T. (2002) Data-driven learning: the perpetual challenge. In Ketteman and Marko (eds.), 107–17.

Johns, T. and King, P. (eds.) (1991) Classroom concordancing. *ELR Journal* 4. University of Birmingham.

Jones, S. and Murphy, M. L. (2005) Using corpora to investigate antonym acquisition. *International Journal of Corpus Linguistics* 10 (3), 401–22.

Kennedy, G. (1998), *Introduction to corpus linguistics*. Harlow, Essex: Longman.

Kennedy, G. (1992) Preferred ways of putting things with implications for language teaching. In Svartvik (ed.), 335–78.

Ketteman, B. and Marko, G. (eds.) (2002) *Teaching and learning by doing corpus analysis*. Proceedings from the Fourth International Conference on Teaching and Language Corpora, Graz July 19–24, 2000. Amsterdam: Rodopi.

Koller, V. and Mautner, G. (2004) Computer applications in critical discourse analysis. In C. Coffin, A. Hewings, and K. O'Halloran (eds.), *Applying English grammar: functional and corpus approaches*. London: Hodder and Stoughton: 216–28.

Krug, M. (2000) *Emerging English modals: a corpus-based study of grammaticalization*. Berlin: Mouton de Gruyter.

Kytö, M. (1996) *Manual to the diachronic part of the Helsinki Corpus of English Texts*, 3rd edn. Helsinki: University of Helsinki Press. http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM

Kytö, M. (1996) "The best and most excellentest way": the rivalling forms of adjective comparison in Late Middle and Early Modern English. In Svartvik (ed.), 123–44.

Kytö, M. and Rissanen, M. (1988) The Helsinki Corpus of English texts: classifying and coding the diachronic part. In Kytö et al. (eds.), 169–79.

Kytö, M., Rissanen, M., and Wright, S. (eds.) (1994) *Corpora across the centuries.* Proceedings of the First International Colloquium on English Diachronic Corpora, St Catharine's College Cambridge, March 22–7, 1993. Amsterdam: Rodopi.

Kytö, M., Ihalainen, O., and Rissanen, M. (eds.) (1988) *Corpus linguistics, hard and soft.* Proceedings of the Eighth International Conference on English Language Research on Computerized Corpora. Amsterdam: Rodopi.

Lancashire, I., Meyer, C., and Carol, P. (eds.) (1996) Papers from English language research on computerized corpora (ICAME 16) Amsterdam: Rodopi.

Landau, I. L. (2001, 2nd edn.) *Dictionaries: the art and craft of lexicography*. Cambridge: Cambridge University Press.

Leech, G. (1991) The state of the art in corpus linguistics. In Aimer and Altenberg (eds.), 8–29.

Leech, G. (1992) Corpora and theories of linguistic performance. In Svartvik (ed.), 105–22.

Leech, G. (1997a) Introducing corpus annotation. In Garside et al. (eds.), 1–18.

Leech, G. (1997b) Teaching and language corpora: a convergence. In Wichmann et al. (eds.), 1–23.

Leech, G. (2003) Modality on the move: the English modal auxiliaries 1961–1992. In Facchineti and Krug (eds.), 223–40.

Lewandowska-Tomaszczyk, B. and Melia, P. J. (eds.) (1997) *Practical Applications in Language Corpora (PALC '97)* Łódź: Łódź University Press.

Lightbown, P. (1985) Can language acquisition be altered by instruction? In Hyltenstam and Pienemann (eds.), 101–12.

Lindquist, H. and Mair, C. (eds.) (2004) *Corpus approaches to grammaticalization in English*. Amsterdam: John Benjamins.

Ljung, M. (ed.) (1997) *Corpus-based studies in English.* Papers from the Seventeenth International Conference on English Language Research on Computerized Corpora (ICAME 17) Amsterdam: Rodopi.

Longman dictionary of contemporary English (2003) 4th edn. Harlow, Essex: Longman.

Lopez-Couso, M. J. and Mendez-Naya, B. (2001) On the history of *if*- and *though*-links with declarative complement clauses. *English Language and Linguistics* 5 (1), 93–107.

Lorenz, G. (1998) Overstatement in advanced learners' writing: Stylistic aspects of adjective intensification. In Granger (ed.), 53–66.

Loschky, L. and Bley-Vroman, R. (1993) Grammar and task-based methodology. In Crookes and Gass (eds.), 123–66.

Louw, B. (1993) Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In Baker et al. (eds.), 157–76.

Macmillan English dictionary for advanced learners (2002) Basingstoke, Hampshire. Macmillan.

McEnery, A. M. (2003) Corpus linguistics. In Mitkov (ed.), 448–63.

McEnery, A. M. (2005) *Swearing in English: bad language, purity and power from 1586 to the present*. London: Routledge.

McEnery, A. M., Wilson, A., and Baker, J. P. (1997) Teaching grammar again after twenty years: corpus-based help for teaching grammar. *ReCALL Journal* 9 (2), 8–17.

McEnery, A. and Kifle, N. (2002) Epistemic modality in argumentative essays of second language writers. In Flowerdew (ed.), 182–95.

McEnery, T. and Wilson, A. (2001) 2nd edn. *Corpus linguistics.* Edinburgh: Edinburgh University Press.

McEnery, A. M. and Xiao, Z. (2004) Swearing in modern British English: the case of *fuck* in the BNC. *Language and Literature* 13 (3), 237–70.

Mair, C. (1991) Quantitative or qualitative corpus analysis? Infinitival complement clause in the Survey of English Usage corpus. In Johansson and Stenström (eds.), 67–80.

Mair, C., Hundt, M., Leech, G., and Smith, N. (2003) Short term diachronic shifts in part-of-speech frequencies: a comparison of the tagged LOB and F-LOB corpora. *International Journal of Corpus Linguistics* 7 (2), 245–64.

Meyer, C. F. (2002) *English corpus linguistics*. Cambridge: Cambridge University Press.

Milton, J. (1998) Exploiting L1 and interlanguage corpora in the design of an electronic language learning and production environment. In Granger (ed.), 186–98.

Mindt, D. (1995) *An empirical grammar of the English verb: modal verbs*. Berlin: Cornelsen Verlag.

Mindt, D. (2000) *An empirical grammar of the English verb system*. Berlin: Cornelsen.

Mitkov, R. (2003) *Handbook of computational linguistics*. Oxford: Oxford University Press.

Nelson, G., Wallis, S., and Aarts, B. (2002) *Exploring natural language: working with the British component of*

the International Corpus of English. Amsterdam: John Benjamins.
Nesselhauf, N. (2005) Collocations in a learner corpus. *Studies in Corpus Linguistics 14*. Amsterdam: John Benjamins.
Nevalainen, T. and Kahlas-Tarkka, L. (eds.) (1997) *To explain the present: studies in the changing English language in honour of Matti Rissanen.* Mémoires de la Société Néophilologique de Helsinki. Helsinki: Société Néophilologique.
Nevalainen, T. and Raumolin-Brunberg, H. (2003) *Historical sociolinguistics*. London: Longman.
Newmeyer, F. J. (2003) Grammar is grammar and usage is usage. *Language* 79 (4), 682–707.
Nunan, D. (1989) *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.
Ooi, V. B. Y. (1998) *Computer corpus lexicography*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.
Oostdijk, N. (1991) *Corpus linguistics and the automatic analysis of English*. Amsterdam: Rodopi.
Orpin, D. (2005) Corpus linguistics and critical discourse analysis: examining the ideology of sleaze. *International Journal of Corpus Linguistics* 10 (1), 37–61.
Osbourne, J. (2000) What can students learn from a corpus? Building bridges between data and explanation. In Burnard and McEnery (eds.), 193–205.
Owen, C. (1993) Corpus-based grammar and the Heineken effect: Lexico-grammatical description for language learners. *Applied Linguistics* 14 (2), 167–87.
Oxford advanced learner's dictionary (2002) 6th edn. Oxford: Oxford University Press.
Oxford dictionary of English (2003) 2nd edn. Oxford: Oxford University Press.
Palmer, F. R. (ed.) (1968) *Selected papers of J.R. Firth 1952–59*. London: Longmans.
Partington, A. (2004) "Utterly content in each other's company": semantic prosody and semantic preference. *International Journal of Corpus Linguistics* 9 (1), 131–56.
Partington, A. Morley, J., and Harman, L. (eds.) (2004) *Corpora and discourse*. Proceedings of CamConf 2002, Università degli Studi di Camerino, Centro Linguistico d'Ateneo, September 27–9. New York: Peter Lang.
Paulillo, J. C. (2000) Formalising formality: an analysis of register variation in Sinhala. *Journal of Linguistics* 36 (2), 215–59.
Phillips, M. (1989) *Lexical structure of text*: *discourse analysis monograph no. 12*. English Language Research, University of Birmingham.
Polovina-Vukovic, D. (204) The representation of social actors in the Globe and Mail during the break-up of the former Yugoslavia. In L. Young and C. Harrison (eds.), *Systemic functional linguistics and critical discourse analysis*. London and New York: Continuum, 155–72.
Prodromou, L. (1997) Corpora: the real thing? *English Teaching Professional* 5, 2–6.
Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985) *A comprehensive grammar of the English language*. London: Longman.
Renouf, A. (1996) The ACRONYM Project: Discovering the textual thesaurus. In Lancashire et al. (eds.), 171–87.
Renouf, A. (1997) Tools for the diachronic study of historical corpora. In Nevalainen and Kahlas-Tarkka (eds.), 185–99.
Renouf, A. (ed.) (1998) *Explorations in corpus linguistics*. Amsterdam: Rodopi.
Renouf, A. (2001) Lexical signals of word relations. In Scott and Thompson (eds.), 35–54.

Rissanen, M. (1997) Introduction. In Rissanen et al. (eds.), 1–15.

Rissanen, M., Kytö, M., and Heikkonen, K. (eds.) (1997) *Grammaticalization at work: studies of long-term developments in English.* Berlin: Mouton de Gruyter.

Römer, U. (2004) Textbooks: a corpus-driven approach to modal auxiliaries and their didactics. In J. McH. Sinclair (ed.), *How to use corpora in language teaching*. Amsterdam: John Benjamins, 185–99.

Römer, U. (2005) *Progressives, patterns, pedagogy: a corpus-driven approach to progressive forms, functions, contexts and dialects*. Amsterdam: John Benjamins.

Sampson, G. (2001) *Empirical linguistics*. London: Continuum.

Schmid, H.-J. (2000) *English abstract nouns as conceptual shells: from corpus to cognition*. Berlin: Mouton de Gruyter.

Schmidt, R. W. (1990) The role of consciousness in second language learning. *Applied Linguistics* 11 (2), 129–58.

Schonefeld, D. (1999) Corpus linguistics and cognitivism. *International Journal of Corpus Linguistics* 4 (1), 137–71.

Scott, M. and Thompson, G. (eds.) (2001) *Patterns of text: in honour of Michael Hoey*. Amsterdam/Philadelphia: John Benjamins.

Seidlhofer, B. (ed.) (2003) *Controversies in applied linguistics*. Oxford: Oxford University Press.

Semino, A. and Short, M. H. (2004) *Corpus stylistics*. London: Longman.

Sharwood Smith, M. (1981) Consciousness-raising and the second language learner. *Applied Linguistics* 2, 159–69.

Simpson, R. C. and Swales, J. M. (eds.) (2001) *Corpus linguistics in North America*. Ann Arbor: University of Michigan Press.

Sinclair, J. McH. (1966) Beginning the Study of Lexis. In Bazell et al. (eds.), 410–31.

Sinclair, J. McH. (ed.) (1987) *Looking up: an account of the COBUILD Project in lexical computing*. London: Collins ELT.

Sinclair, J. McH. (1991) *Corpus concordance collocation*. Oxford: Oxford University Press.

Sinclair, J. McH. (1998) The lexical item. In Weigang (ed.), 1–24.

Sinclair, J. (2004a) Intuition and annotation: the discussion continues. In Aijmer and Altenberg (eds.), 39–59.

Sinclair, J. (2004b) *Trust the text: language, corpus and discourse*. London: Routledge.

Sinclair, J. McH. (ed.) (2004c) *How to use corpora in language teaching*. Amsterdam: John Benjamins.

Skehan, P. (1998) *A cognitive approach to language learning*. Oxford: Oxford University Press.

Smith, N. (2003) Changes in the modals and semi-modals of strong obligation and epistemic necessity in recent British English. In Facchineti and Krug (eds.), 241–66.

Sotillo, S. M. and Wang-Gempp, J. (2004) Using corpus linguistics to investigate class, ideology and discursive practices in online political discussions: pedagogical applications of corpora. In U. Conner and T. A. Upton (eds.), *Applied corpus linguistics*. Amsterdam: Rodopi, 91–122.

Stefanowitsch, A. (2005) The function of metaphor: developing a corpus-based perspective. *International Journal of Corpus Linguistics* 10 (2), 161–98.

Stubbs, M. (1996) *Text and corpus analysis: computer-assisted studies of language and culture.* Oxford: Blackwell.

Stubbs, M. (2001) *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.

Stubbs, M. (2002) Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7 (2), 215–44.

Stubbs, M. (2005) Conrad in the computer: examples of quantitative

stylistic methods. *Language and Literature* 14 (1), 5–24.
Svartvik, J. (1966) *On voice in the English verb*. The Hague: Mouton and Co.
Svartvik, J. (ed.) (1992) *Directions in corpus linguistics: proceedings of the Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin: Mouton de Gruyter.
Svartvik, J. (ed.) (1996) *Words: proceedings of an international symposium, Lund, 25–26 August 1995.* Stockholm: Kungl. Vitterhets Historie och Antikvitets Akademien.
Teubert, W. (1999) Corpus linguistics: a partisan view. *TELRI Newsletter* April (8), 4–19.
Tognini-Bonelli, E. (2001) *Corpus linguistics at work*. Amsterdam: John Benjamins Publishing Company.
Tono, Y. (2000) A computer learner corpus based analysis of the acquisition order of English grammatical morphemes. In Burnard and McEnery (eds.), 123–32.
Tribble, C. (1997) Put a corpus in your classroom: using a computer in vocabulary development. In Boswood (ed.), 266–8.
Tribble, C. and Jones, G. (1990) *Concordances in the classroom*. London: London Group UK Limited.
Virtanen, T. (1997) The progressive in NS and NNS student compositions: evidence from the International Corpus of Learner English. In Ljung (ed.), 299–309.
Virtanen, T. (1998) Direct questions in argumentative student writing. In Granger (ed.), 94–106.
Vivanco, V. (2005) The absence of connectives and the maintenance of coherence in publicity texts. *Journal of Pragmatics* 37 (8), 1233–49.
Wang, S. (2005) Corpus-based approaches and discourse analysis in relation to reduplication and repetition. *Journal of Pragmatics* 37 (4), 505–40.
Weigang, E. (ed.) (1998) *Contrastive lexical semantics*. Amsterdam and Philadelphia: John Benjamins Publishing Company.
Wichmann, A., Fligelstone, S., McEnery, T., and Knowles, G. (1997) *Teaching and language corpora.* New York: Addison Wesley Longman.
Widdowson, H. G. (1991) The description and prescription of language. In Alatis (ed.), 11–24.