Moon R (2001). 'The distribution of idioms in English.' *Studi Italiani di Linguistica Teorica e Applicata 2001/2,* 229–241.
Newmeyer F J (1972). 'The insertion of idioms.' In Peranteau P M, Levi J N & Phares G C (eds.) *Papers from the eighth regional meeting of the Chicago Linguistic Society.* 294–302.
Pulman S (1993). 'The recognition and interpretation of idioms.' In Cacciari C & Tabossi P (eds.) *Idioms: processing, structure, and interpretation.* Hillsdale, NJ: Lawrence Erlbaum Associates. 249–70.
Sinclair J (1998). 'The lexical item.' In Weigand E (ed.) *Contrastive lexical semantics.* Amsterdam, Philadelphia: John Benjamins. 1–24.
Stubbs M (2002). 'Two quantitative methods of studying phraseology in English.' *International Journal of Corpus Linguistics 7(2),* 215–244.

Corpus Linguistics

S Hunston, University of Birmingham, Birmingham, UK

Introduction

A corpus is an electronically stored collection of samples of naturally occurring language. Most modern corpora are at least 1 million words in size and consist either of complete texts or of large extracts from long texts. Usually the texts are selected to represent a type of communication or a variety of language; for example, a corpus may be compiled to represent the English used in history textbooks, or Canadian French, or Internet discussions of genetic modification. Corpora are investigated through the use of dedicated software.

Corpus linguistics can be regarded as a sophisticated method of finding answers to the kinds of questions linguists have always asked. A large corpus can be a test bed for hypotheses and can be used to add a quantitative dimension to many linguistic studies. It is also true, however, that corpus software presents the researcher with language in a form that is not normally encountered and that this can highlight patterning that often goes unnoticed. Corpus linguistics has also, therefore, led to a reassessment of what language is like.

This article discusses the resources and methodologies used by corpus linguists and then moves on to some key observations relating to comparative frequency and to patterning. It also considers the importance of corpus linguistics for linguistic theory and presents some of the applications of corpus research.

Resources and Methodologies for Corpus Linguistics

Corpora

The basic resource for corpus linguistics is a collection of texts, called a corpus. Corpora can be of varying sizes, are compiled for different purposes, and are composed of texts of different types. All corpora are homogeneous to a certain extent; they are composed of texts from one language or one variety of a language or one register, etc. They also are all heterogeneous to a certain extent, in that at the very least they are composed of a number of different texts. Most corpora contain information in addition to the texts that make them up, such as information about the texts themselves, part-of-speech tags for each word, and parsing information.

Searches, Software, and Methodologies

Corpora are interrogated through the use of dedicated software, the nature of which inevitably reflects assumptions about methodology in corpus investigation. At the most basic level, corpus software:

- searches the corpus for a given target item,
- counts the number of instances of the target item in the corpus and calculates relative frequencies,
- displays instances of the target item so that the corpus user can carry out further investigation.

It is apparent that corpus methodologies are essentially quantitative. Indeed, corpus linguistics has been criticized for allowing only the observation of relative quantity and for failing to expand the explanatory power of linguistic theory (for discussion, see Meyer, 2002: 2–5). It is shown in this article that corpus linguistics can indeed enrich language theory, though only if preconceptions about what that theory consists of are allowed to change. Here, however, we leave that argument aside as we review corpus investigation software in more detail.

Search Items

Any corpus, including a 'raw' corpus (that is, a corpus that consists only of text, with no further information added), may be searched for instances of a single word (e.g., *day*). Most search software

also allows a single search to find sets of words (e.g., *day*, *month*, *year*) and strings of words (e.g., *the next day*).

If the corpus is tagged for part of speech, the tags can form part of the search. Depending on the software, it is possible to search for a word when it is tagged for a particular word class only, such as *light* when it is tagged as an adjective, not as a noun or a verb; given sequences of tags, such as 'preposition + determiner + noun'; or individual words followed or preceded by given word classes, such as *fundamentally* followed by an adjective. Similarly, a corpus that is parsed will allow searches for particular clause types or structures. For example, Nelson *et al.* (2002) describe searching the International Corpus of English for sentences containing an *if* clause before or after the main clause.

Finally, corpora can be annotated for other kinds of information, such as semantic categories, categories of cohesion, or the representation of speech and thought (Garside *et al.*, 1997). Software can count the occurrence of such categories and, usually, compare their frequency in corpora of different kinds.

Word Lists and Frequency Information

A word list is a list, usually arranged either alphabetically or in frequency order, of all the words in a given corpus with information about the number of times that word occurs in the corpus. The simplest word lists interpret 'word' as simply a string of letters; so, for example, the number of occurrences of *run* is given without distinction between the noun and the verb, and the occurrences of *runs, running*, and *ran* are given separately. More sophisticated lists distinguish between, say, the noun and verb occurrences of *run* and give summary figures for a whole lemma, such as for *run, runs, running, ran*, all occurring as verbs (Leech *et al.*, 2001). Much more difficult, and indeed not publicly available, are word lists that distinguish between senses (e.g., between *run* meaning 'move in fast motion' and *run* meaning 'manage an event or organization').

Comparative Frequencies

Information about frequency is not very informative unless it is comparative. In the Nelson *et al.* (2000) study of *if* clauses, for example, it is noted that whereas in written registers and formal spoken registers these clauses are more frequent before the main clause than after, in informal spoken registers the reverse is the case. Thus, information about frequency is generally used to compare one corpus with another and, by implication, to compare two languages, varieties of a language, or text types.

Unless the two corpora being compared are exactly the same size, raw frequency information is of limited use. For example, the word *quite* occurs 11 441 times in a corpus consisting of issues of the *Times* (London) newspaper and 22 594 times in a corpus consisting of spoken British English. That is, there are about twice as many instances in the spoken corpus as in the *Times* corpus. However, the two corpora are very different in size. (Both corpora are part of the Bank of English corpus, owned jointly by the University of Birmingham and the publisher HarperCollins. The *Times* corpus contains about 51 million words, and the spoken corpus about 20 million.) It is more helpful to say how many times *quite* occurs, on average, in every million words of each corpus. This shows that *quite* is even more frequent, relatively, in the smaller, spoken corpus than the raw figures suggest. It occurs 1125 times every million words in the spoken corpus and 220 times per million words in the *Times* corpus. In other words, it is used about five times as frequently in the spoken corpus as in the *Times* corpus.

What is still uncertain from a calculation such as this one is how important such a difference is. Is the higher incidence of *quite* in the spoken corpus sufficient to consider it to be a marker of speech? Log likelihood calculations are often used to rank differences between corpora. According to Leech *et al.* (2001), *quite* is about the hundredth most different word in a comparison between the spoken and written components of the British National Corpus, with a very high log likelihood of about 8500.

A similar calculation is used to compare relatively small corpora of specialized texts with larger, more general corpora, using the Keywords program (part of the Wordsmith Tools suite, Scott, 1996). Keywords ranks the words in the specialized corpus in order of the magnitude of their difference from the general corpus. This indicates what makes the specialized texts different from English in general. Most Keywords are lexical words that reflect the specific content (or the 'aboutness,' in Scott's terms) of the text or small corpus. For example, a corpus of newspaper feature articles, when compared with a more general corpus of newspaper texts, is found to have Keywords such as *tax*, *European*, *war*, *education*, *schools*, and *church* (Scott, 2001: 116). These reflect the prevalent themes of the articles in question. Some Keywords, more surprisingly, are grammatical items, such as pronouns, prepositions, or *be* (Scott, 2001: 126). Such words often occur in specific phraseological sequences that are more frequent in the specialized corpus than in the general one. Gledhill (2000), for example, notes that sections of research articles can be distinguished in this way.

Collocations

Collocation is the tendency of some words to co-occur more frequently than others. The words that frequently cooccur with a node word are known as the collocates of that node word. Software that calculates the frequency of collocates has to work on a given 'span,' that is, a set number of words before and after the node. A span of 4:4 is often used (that is, four words before the node and four words after it), but most software allows the user to specify the span. The software then gives a list of the words that occur within the selected span.

The simplest collocate list is in order of raw frequency, but this tends to include words that are not particularly significant for a given node word but that are very frequent in the language as a whole. In English, for example, *the* tends to occur near the top of many collocate lists simply because it is so frequent overall. Statistical packages (such as *t*-score, *z*-score, or mutual information) are often used to correct for this. These calculations compare the actual number of occurrences of a given word as a collocate with the number of occurrences that would be expected if the words in the corpus were distributed randomly. Depending on the package used, a list of collocates ordered according to significance can prioritize words that are unusual outside the context of the node word (such as *unblinking gaze*, where *unblinking* is infrequent except when it occurs with *gaze*) or words that have a wider range of behaviors (such as *his or her gaze*, where the possessives of course occur with a multitude of nouns – it is, however, significant that *gaze* is often preceded by a possessive) (Hunston, 2002: 69–74).

Collocates can be shown as a simple list, or the list can be organized to show where in relation to the node word each collocate normally appears. For example, Table 1 shows a list of the 15 most frequent collocates of *gaze* (Hunston, 2002: 69–70).

The 10 most significant collocates occurring immediately before *gaze* are *his*, *her*, *'s*, *my*, *to*, *public*, *their*, *your*, *our*, *steady*. The 10 most significant collocates in the position immediately after *gaze* are *at*, *of*, *on*, *into*, *upon*, *out*, *from*, *was*, *and*, *fixed*. Significant collocates occurring two places to the left of *gaze* include *under*, *turned*, *followed*, *shifted*, *avert*, *fixed*, and *returned*.

Table 1 Collocates of *Gaze*

the	1511	he	277	under	154
his	822	from	228	their	140
her	628	with	225	public	109
's	442	she	213	fixed	102
on	333	my	177	then	86

Concordance Lines

All the software discussed so far carries out a number of statistical operations on items found in the corpus, ranging from simply counting the number of occurrences to measuring the degree of significance of occurrence. In contrast, software that presents concordance lines simply identifies the target item (usually a word or phrase) each time it occurs in the corpus and presents each instance, or as many as are required, to the corpus user. Usually this is done with the target item in the center of the screen and a few words to the left and right of that item. This 'key word in context' presentation, as it is known, has a number of uses. Even the small amount of context is usually enough to show what the word or phrase means, what phrases it often occurs in, and/or the discourse function that it has. Quantitative information about word meaning and function that is not available automatically can therefore be calculated.

Importantly, concordance lines can usually be sorted so that the word(s) coming before or after the node word are arranged alphabetically. This has the advantage of making more obvious the recurring phraseology that many words are part of and that are not revealed by lists of collocates.

Concordance lines are usually of a length so that they will fit neatly in a computer window or on a normal size piece of paper, with each 'line,' or instance of the target word, occupying one line of print (see Figure 1). This makes patterning easy to observe. It is, however, only a convenience, and often users find it beneficial to look at more context – that is, to look at concordance lines that are much longer than the width of a computer window. Most software will allow the concordance line to be expanded in this way.

Comparing Frequency of Occurrence

Most of the software mentioned thus far in this article is essentially comparative. Word lists and collocation software compare the frequencies of two items in a single corpus. Log likelihood and Keywords compare the frequency of items in more than one corpus. Comparative frequency is the essence of many kinds of corpus research.

Types of Comparisons Made

Comparisons are often made between different languages or varieties, between different registers or types of text, and between different historical periods.

Languages and Varieties

Studies that compare different languages use two kinds of corpora, here given the labels 'parallel corpora' (where one corpus consists of translations of texts in the other corpus) and 'comparable corpora' (where each corpus consists of the same kind of texts written originally in each language). These studies, which can show the difference in frequency of particular features, are often used to demonstrate the lack of direct equivalence between apparently similar aspects of related languages. For example, Johansson (1996) notes that although *wh-* cleft sentences, such as "What we need is a new car," occur in both English and Swedish, they are more frequent in English than in Swedish. Translations from English into Swedish tend to replace the *wh-* cleft construction with an alternative word order, while translations from Swedish into English only occasionally replace the alternative with a *wh-* cleft. As a result, texts that are English translations of Swedish contain an unusually low number of *wh-* clefts. Another example is the English word *and* and its translation equivalents in French (*et*) and Swedish (*og*). Comparable corpora that contain equivalent original texts of the three languages show that *et* is less frequent in French than *and* is in English and that *og* is more frequent in Swedish than *and* is in English (Allén, 1970). Comparing individual translated examples can be very revealing of how these differences in frequency come about. For example, Schmied and Fink (2000) notes that in their corpus the English word *with* is translated using the apparent German equivalent *mit* in fewer than half the instances. Translators sometimes use other prepositions (e.g., *with a city as compact as Cardiff* translated as *bei einer kompakten stadt wie Cardiff*), or other parts of speech (e.g., *with energy* translated using the adverb *energisch*) or choose an alternative way of expressing the idea (e.g., *the advent of secondary school for all children, with more democratic access to university* is translated with *und* 'and' joining the two noun phrases, instead of *mit*).

The work on collocation and meaning has also been applied crosslinguistically. Words that appear to be synonyms, because they are translated by a single item in another language, may be shown to have different collocates. Teubert and Čermáková (2004: 153), for example, show that while the French word *travail/travaux* can be translated as either *work* or *labor* in English, the collocates of *work* and *labor*, when these translate *travail*, do not overlap. In other words, although *work* and *labor* have the same translation into French, they are used in English in completely different contexts.

Varieties of a single language are also compared. Meyer (2002: 125), for example, compares the frequency of 'pseudo-titles' (e.g., *lawyer Freda Jones*) in newspapers from different countries where English is a major language. He confirmed what has long been known, that newspapers in the United States use such titles more than those in Britain. It might be expected, then, that the Philippines, influenced by America, would use such titles more than New Zealand, influenced by Britain. In fact, Meyer finds that pseudo-titles are more frequent in both the Philippines and the New Zealand press than they are even in the U.S. press.

U.S. and British English were extensively compared in Biber *et al*. (1999). Among their findings were the following:

- Progressive aspect is used much more frequently in American conversation than in British conversation, whereas the perfect aspect is used more frequently in British news reporting than in the American equivalent (462).
- The modals *will*, *(had) better*, and *(have) got to* appear more frequently in British conversation than in American, whereas *have to* and *be going to* appear more frequently in American conversation (488).
- British conversation uses more initial and final ellipses than American conversation does, whereas American conversation uses more medial ellipses than British conversation does (1108).
- Negative interrogatives were used about twice as often in British conversation as in American (1115).

Register

Much of the comparative work using corpora has compared a language such as English as it appears in different contexts. These contexts have sometimes been defined in line with a linguistic theory (e.g., in Matthiessen, 2005, where register is defined according to systemic theory), or according to a less theoretical, 'commonsense' view of where clear distinctions might lie. Biber *et al*. (1999), for example, take broad 'register' categories of conversation, fiction, news reporting, and academic prose. Others have made more refined distinctions: the CANCODE corpus of spoken English, for example, distinguishes between 'transactional,' 'professional,' 'socializing,' and 'intimate' contexts (Carter, 2004); Hyland

(2000) distinguishes between academic genres such as research articles, book reviews, abstracts, and textbooks and between different academic disciplines. Much of the work involved in register comparison has focused on explaining quantitative results qualitatively. Mostly, variation in frequency has been accounted for in terms of differences in communicative function or rhetorical purpose.

Comparisons between registers may focus on single words, or more usually on sets of words, phrases, or patterns that are known to share a meaning or function, with the aim of describing variation between registers in the realization of that function. Conrad and Biber (2000), for example, compare frequencies of stance adverbials in conversation, news reporting, and academic prose. They find that such adverbials are most frequent in conversation and that single adverbs expressing epistemic meaning are the most frequent, though prepositional phrases are relatively frequent in academic prose and news. Hyland (2001) compare the frequency of self-citation (where an author refers to other papers he or she wrote) in eight different academic disciplines, including physics, marketing, sociology, and mechanical engineering. He notes that self-citation is more frequent in sciences than in humanities or social sciences. Hyland argues that this difference can be accounted for by the greater reliance in science on a given body of previous work in a relatively small disciplinary area, whereas research in humanities tends to be drawn from a wider base and does not rely as much on immediately preceding research. Semino and Short (2004) take categories of representation of speech and writing (such as direct speech and free indirect speech) and compare their frequency in corpora compose of fiction, newspapers, and (auto)biography. They find that direct speech is the most frequent category overall and that it is particularly frequent in fiction. Newspapers use more indirect speech, and summaries of speech events, than fiction does. (Auto)biography comes between the two other corpora, in that it uses less direct speech than fiction and less indirect speech than newspapers.

Register comparisons focusing on grammatical categories are also common. The largest of these is Biber *et al.* (1999), which compares the frequency of almost all grammatical categories between conversation, fiction, news reporting, and academic prose. Among the categories are word class (more nouns in the written registers, more pronouns and verbs in conversation), clause types (there were more interrogative and imperative clauses in conversation than in the other registers), and tense and aspect (more present tense in conversation and academic prose, more past tense in fiction, about equal proportions in news).

Historical Period

Comparisons between historical periods are difficult because of the lack of truly comparable corpora. Although substantial numbers of texts from earlier periods are now available electronically, there is nowhere near the same quantity or variation as is available for contemporary texts. However, some work is possible, and this can show how both language and ideas have changed over time. Using the multidimensional approach described in this article's section 'Multidimensional Variation' for example, Biber and Finegan (2001) show that between 1650 and 1990 the conversations represented by a corpus of drama became stylistically closer to modern casual conversation, demonstrating 'involvement,' a personal style, and a close connection to the immediate situation. Medical prose, on the other hand, over the same time period became less involved, more impersonal, and more removed from the immediate context. Whereas in 1650, drama and medical prose were relatively close to each other stylistically, over the centuries they became more distinct in style, with drama adopting the characteristics that we now associate with speech, and medical prose adopting those associated with writing.

Using the same time period but a different methodology, Hundt (2004) traces the decline of the 'passival' (active clauses with passive meaning, such as *The house is building*) and the rise of the 'progressive passive' (e.g., *The house is being built*). In the part of the corpus dating from 1650 to 1800, only the passival was found. The progressive passive first appeared in the early 19th century, but it then increased rapidly in frequency, becoming far more frequent, relatively, than the passival ever was. By 1990 the passival had all but disappeared.

It is not only style and grammar that can be shown to change over time. Teubert (2004b) traces the development of the English word *guilt* through a succession of literary and other texts, including those by Shakespeare, Milton, and George Eliot. He shows that although current writers have often ascribed feelings of guilt to characters in those works, in fact *guilt* was not used as a word to describe feelings until the second half of the 19th century. He further argues that this represents a change in the collective consciousness in the English-speaking world at that

time and that the change in both language and consciousness is attributable to the work of some German writers, including Freud.

Issues in Comparison

A number of themes recur in discussions of comparison between registers and deserve special mention: the difference between speech and writing; multidimensional variation; and probability.

Speech and Writing The variation between spoken and written English has been studied through the use of corpora by, among others, McCarthy and Carter (2002) and Biber *et al.* (1999). Their findings can be summarized thus:

- There is a high incidence in spoken English of hesitations, false starts, repetition, and other features attributable to the difficulty of producing language quickly and spontaneously. These features are missing in written English.
- There is a difference in the frequency of some grammatical categories, attributable to various aspects of the contexts of speech and writing, such as the greater possibility of exophoric reference in face-to-face communication, which accounts for the higher incidence of pronouns in speech.
- There is a difference in the frequency of some semantic or pragmatic categories, such as the higher incidence of 'vague language' in speech, attributable to the affective and interpersonal qualities of speech.
- There are present in speech certain grammatical features that occur regularly and without speaker comment, which are not dialectal and yet which occur rarely in written English. Some of these would be considered 'ungrammatical' in formal written English. These include the following:

1. Ellipsis of the subject, or the subject and operator, where this is not recoverable from context ('situational ellipsis') – e.g., *Don't know* (for 'I don't know') or *Going clubbing tonight?* (for 'Are you going clubbing tonight?')
2. Occurrence of a noun phrase preceding or following a clause that repeats a clause element such as the subject or object – e.g., *That friend of yours, he's quite nice* or *He's quite nice, that friend of yours*
3. Occurrence of freestanding clauses beginning with a subordinator, such as *which*, and of freestanding nonclausal phrases, such as *Nice day*
4. Occurrence of verbs in progressive aspect that in written English are rarely or never progressive – e.g., *I was saying* or *I was looking*.

The question arises as to whether these features suggest that spoken and written English are different varieties warranting different descriptive categories or whether a single grammar will account for both, with differences in frequency and contextual constraints duly noted.

McCarthy and Carter (2002: 51) argue that "spoken grammars have uniquely special qualities that distinguish them from written ones." Carter and McCarthy (1995: 141) make the less radical claim that grammar based on written English only does not describe patterns that nonetheless recur in spoken English. Carter (2004: 57) describes speech and writing as a continuum. In all, they seem to suggest that the two modes share a good deal of common ground, but that some aspects of each need to be described separately, as a description based on one would not apply to the other.

Multidimensional Variation The notion of multidimensional variation was developed by Biber (1988). Biber noted that statistically features tend to cluster in texts, so the presence of one feature makes the presence of given other features more likely and the presence of others less likely. For example, present tense verbs tend to be positively associated with pronouns, *be* as main verb, contractions, hedges, and amplifiers and to be negatively associated with nouns, prepositions, attributive adjectives, and long words (Conrad and Biber, 2001: 23). Having identified this bundle of cooccurring features, Biber (1988) glosses the meaning of a high level of incidence of the features as 'involved,' while a low incidence is 'informational.' He then maps a number of registers (that is, corpora consisting of texts of the same type) onto this dimension. Informal spoken registers, such as conversations, appear high on the dimension – that is, they are highly 'involved.' Formal written registers, such as academic prose and official documents, appear low – that is, they are highly 'informational.' There is, however, considerable overlap between written and spoken registers: personal letters and interviews appear close to each other on the dimension, as do romance fiction and prepared speeches, professional letters and broadcasts. Biber identifies seven dimensions in all, although most work in this area highlights dimensions 1 and 2 (Conrad and Biber eds., 2001):

1. Involved – informational
2. Narrative – nonnarrative
3. Elaborated reference – situation-dependent reference
4. Overt expression of argumentation
5. Abstract style – nonabstract style

6. Online informational elaboration marking stance
7. Academic hedging.

What is most interesting about Biber's work is the observation that registers are not consistently like or unlike each other. For example, academic prose and press reportage score exactly the same on dimension 1, but on dimension 2 academic prose is strongly negative, while press reportage is slightly positive; on dimension 3 academic prose is strongly positive and press reportage is slightly negative. Work by Biber, Conrad, and others has used this multidimensional approach to compare corpora of texts from different historical periods, academic texts from different disciplines, and corpora of British and American English (Conrad and Biber, 2001).

Probability Halliday (1978) describes language as a semiotic system consisting, in abstract, of a series of interlocking networks. The networks represent series of choices that can be made in the language (for example, between interrogative, imperative, and declaration moods, or between present and past tenses, or between mental, material, and relational processes). Each choice can be assigned a probability of occurrence based on its actual frequency of occurrence in a large corpus. For example, Halliday and James (1993) calculate that in a large general corpus (the Bank of English), present and past tenses occur with approximate equal frequency, while positive clauses occur overwhelmingly more frequently than negative ones, in a proportion of about 9:1. A further stage in the argument, however, is that each register – that is, the discourse found in each contextual configuration – has a set of frequencies, and therefore of probabilities, that are different from the general overall probability. More accurately, the overall probability is an amalgamation of the probabilities of all the registers making up the whole. Furthermore, the probabilities of occurrence in each register are themselves the outcome of the frequency of a given feature in each of a number of texts.

Matthiessen (2005) illustrates this concept and notes that whereas some probabilities remain fairly constant across texts and across registers, others vary considerably. For example, comparing written news reports with spoken interviews, he notes that one news report that he studies uses almost 100% positive clauses, whereas one interview uses about 93% positive. The news reports as a whole use about 96% positive, showing the one sample text to be a fairly extreme example. This incidence of positive clauses is just a little higher than it is in all the written texts in Matthiessen's corpus (95%). The interviews taken together use about 90% positive, which is about the same as all the spoken texts. The overall proportion of positive to negative in Matthiessen's corpus is just over 9:1 – that is, about the same as Halliday's calculation for the Bank of English – and the most extreme variation, between all interviews at 90% positive and one news report at 99% positive, is not great. For process types, however, the differences are more pronounced. With the whole corpus taken together, material and relational processes occur with equal frequency. In all the spoken texts taken together, they also occur with nearly equal frequency, but in all the written texts taken together, material processes occur in more than 60% of clauses, whereas relational processes occur in fewer than 40%. In news reports, both material and relational processes occur proportionally less frequently than in written texts as a whole, the difference being accounted for by the greater frequency of verbal processes. Perhaps most interesting are the interview texts, which overall use a lower proportion of material processes and a greater proportion of relational processes than all spoken texts do. The one sample interview text highlighted by Matthiessen, however, does not follow this trend. In this text, the proportion of material processes is higher than the proportion of relational processes. In fact, the proportion of material processes in the sample interview text is about the same as that in the sample news report text.

Systemic linguistics emphasizes the interrelation between a single instance of language – a single text – and the register to which it belongs, and between that register and the language as a whole. Individual choices made in a single text affect the frequencies in the register and in the language, but at the same time those choices acquire meaning from their alignment with or difference from the norm for the register and for the language. Because probability and register are essential components of systemic linguistic theory, the assessment of frequency of the kind carried out by Matthiessen is an essential component of that theory.

Observing Patterned Behavior

Phraseology

Observing Repetition in Concordance Lines Concordances lines bring together and display instances of use of particular word from the widely disparate contexts in which it occurs. When concordance lines are sorted alphabetically, repetitions can be seen that are not obvious when the word is encountered in the ordinary course of reading or listening. The repetitions can be of individual words or of types of words. For example, Figure 1 shows 30 concordance lines for the word *fact*, randomly selected, using

```
 1 ctions, according to him it remained a fact that the interactions tended t
 2 that I am not/do not look Caucasian, a fact that seems to have led the gue
 3  makes it not as simple as it seems, a fact that can discourage the housek
 4 o present what is taken-for-granted as fact, the presence of items disting
 5 chieved much better results and may in fact be able to read English reason
 6 sals: a number of students had not, in fact, been aware that faculty have
 7 s of the two groups of judges were, in fact, comparable, or whether the t
 8 o modify their utterances by hedges in fact demonstrates a scholarly order
 9  of text which is located close by: in fact, following immediately. On the
10 be no harsher than the ESL raters, in fact, if anything, the reverse. T
11  adequate.  Representativeness is, in fact, often regarded as being in di
12 r, we discovered that students are in fact present as architects of their
13 than in the other one Bell taught. In fact, there were so many questions
14 he highest in demand (Tsui, 1992); in fact, this study concluded that uni
15 cieties. What she was arguing for, in fact, was NA beyond the workplace.
16 cal production should not conceal the fact that  scientific production wo
17 draw the attention of students to the fact that their language choices r
18 the early 19th century was due to the fact that 19th century  Europe alr
19 issue is complicated, however, by the fact that our EAP classes are freq
20 mple B (real world).  In spite of the fact that flotation processes hav
21 rent perceptions of tasks reflect the fact that all views of them are so
22 est of clinical medicine. Despite the fact that the participants achieved
23 ch a possibility was indicated by the fact that strikingly  well-preserve
24 l pronouns, particularly I or we  The fact that the writer of a single-au
25  the humanities. On the contrary, the fact that two research students in
26 tion, his use of power over them. The fact that the students pay tuition,
27 nning of the talk what effect did the fact that the vehicle was a Japa
28 g from the students' feedback was the fact that, despite  their diverse n
29(Swales,1990). This was related to the fact that, at that time, human agen
30 ate directly to client policies. This fact may have an influence over the
```

Figure 1 Instances of the word *fact*.

Wordsmith Tools, from the 350 instances of this word in a corpus of research articles in the field of applied linguistics. The lines have been sorted so that the words preceding and following *fact* are in alphabetical order.

It is immediately apparent that *in fact* and *the fact that* occur very frequently. Together these phrases account for 25 out of the 30 lines. To investigate further the behavior of *the fact that*, a random 24 lines from the same corpus are shown in Figure 2.

These lines reveal more repetitions: *due to the fact that* is relatively frequent, as are *given the fact that* and *reflect(s) the fact that*. More generally, the lines indicate that *the fact that* typically follows phrases that do one of two things:

- They indicate that a fact is or is not noticed: *take into account*; *take into consideration*; *conceal*; *neglect*; *points to*; *draw attention to* (lines 1, 3, 4, 12, 23, 24).
- They indicate the relationship between one fact and another: this relationship is often one of contrast (*despite*; *regardless of*), or cause (*reflect*; *attributed to*; *due to*), but other relationships are found (*compounded by*; *independently of*; *given*; *in addition to*) (lines 2, 6, 8, 9, 13, 14, 15, 16, 17, 19, 20, 21, 22). Lines 10 and 11, which have to be expanded to show more context before sense can be made of them, are other examples of 'addition': *An additional consideration . . . is the fact that . . . Another interesting issue . . . is the fact that. . .*

Phrases and Phraseology The discussion of *fact* illustrates the difference between phrases and phraseology. *In fact* is an example of a 'fixed phrase,' in that it always occurs in the same form and behaves like a single word. *The fact that* is a frequently recurring phrase that is nonetheless not fixed. *A fact that* is found, though it is not as frequent; occasionally other words interrupt the sequence, as in *the very fact that*. Also, other words can replace *fact*, giving sequences such as *the assumption that*, *the idea that*, *the notion that*, and *the view that* (all examples come from the same corpus as in the previous section). On the other hand, sequences such as *due to the fact that* are examples of the broader phraseological behavior of *fact*, which includes nonrecurring items that nevertheless share a type of meaning. In other words, even though the sequence *in addition to the fact that* occurs only once in the corpus mentioned in the previous section, and therefore cannot possibly constitute a phrase, *in addition to* is still part of the phraseology of *fact*.

There is a considerable amount of research that investigates the occurrence of phrases in corpora. 'Phraseology,' meaning the investigation of phrases, is a recognized topic in lexicography, for example, and has been invigorated by the introduction of corpus techniques, which make the identification of phrases more accurate (Cowie, 1998).

The study of phraseology, the tendency for words to occur in some environments more than others, is more recent and also arises from lexicography.

```
 1 However, if we take into account the fact that the students were studying
 2  schemata. This is compounded by the fact that the texts should ideally be
 3 al production should not conceal the fact that scientific production worl
 4 cially taking into consideration the fact that the `Bulgarian deviations'
 5 the humanities. On the contrary, the fact that two research students in su
 6 st of clinical medicine. Despite the fact that the participants achieved v
 7 ning of the talk what effect did the fact that the vehicle was a Japanese
 8 they and their peers make. Given the fact that we are dealing with academ
 9 th effective language use. Given the fact that English is the predominant
10 ing these exit-level courses is the fact that most students who have rea
11 brought about in the results is the fact that some choices can perform mo
12 hat this position also neglects the fact that most ESP/EAP learners need
13 uns and verbs, independently of the fact that, as previously pointed out,
14 onal interaction, regardless of the fact that it can be stored, retrieved
15 nt perceptions of tasks reflect the fact that all views of them are socia
16 ation across faculties reflects the fact that some of them are more favou
17 e (consider this in addition to the fact that these domain experts are ES
18 ten source can be attributed to the fact that scientific activity was—and
19 e early 19th century was due to the fact that 19th century  Europe alread
20    This may be partially due to the fact that Lecturer A had required gro
21 mdemn loss of body water due to the fact that the the loss of body water
22 end-list patterns can be due to the fact that the number of printed pages
23 then, this literature points to the fact that different disciplines ident
24 aw the attention of students to the fact that their language choices refl
```

Figure 2 Instances of the *fact that*.

Sinclair (2004) focuses on sequences that are not fixed but that show the kinds of recurrence demonstrated in the discussion of *the fact that* in the previous section. He proposes the concepts of 'core,' 'collocation,' 'colligation,' and 'semantic prosody' to account for these. Taking the example of the verb *budge*, he notes that it is used in sequences such as *refuse to budge*, *could not budge me from*, *I determined not to budge from it*, *two horses could not budge it*, *they won't budge from that position*, and so on. These are not phrases, as such, yet they clearly have a lot in common. Sinclair describes the behavior of *budge* as shown in Table 2.

Stubbs (2001: 87–88) renames 'semantic prosody' 'discourse prosody' and uses the term 'semantic preference' to indicate collocation with a set of semantically related items. As an example, he discusses the word *undergo*, which is used in examples such as *he was forced to undergo an emergency operation*; *his character appeared to undergo a major transformation*; *each operative had to undergo the most rigorous test*; *forced to . . . undergo further migration and further suffering* (Stubbs, 2001: 92). Stubbs describes the behavior of the word as shown in Table 3.

An additional point, made by Louw (1993), is that the semantic or discourse prosody of a word may be exploited to imply a meaning that is not stated outright. For example, an utterance such as 'To keep my family happy I had to undergo three weeks' holiday by the Mediterranean' implies that a holiday, for this speaker, is something unpleasant. A hearer might interpret this as irony, or humor, or sheer eccentricity.

Pattern and Meaning

Phraseology and Meaning One of the key points about phraseology is that it is closely connected to meaning. Corpus-driven lexicography has indicated that where a word has two or more distinct meanings, each will tend to occur in a specific phraseology. Sinclair (2003) gives several examples, one of which is *block*, a word that can be a noun or a verb. It has a number of meanings, but it is very rarely ambiguous, because the meaning is identifiable from the immediately surrounding phraseology:

- A building: *cell block, administration block*
- A group of buildings between two streets: *half a block, down the block*
- Something hindering someone: *stumbling block*
- A large piece of solid material: *block of stone*
- Stop someone or something from progressing: *block enemy penetrations*, *block the chemical signals*, *block such a move*
- Create an obstacle: *block your path*; *block your own good*.

Often, difference in meaning correlates with grammatical pattern. For example, the verb *maintain* has two main meanings: 'keep something in good condition' and 'say that something is true.' In the first meaning, the verb is usually followed by a noun phrase (e.g., *maintain a road*, *maintain a friendship*), while in the second meaning, it is usually followed by a *that* clause (e.g., *she maintained that . . .*).

Pattern Grammar A specific application of the observation that phraseology and meaning are linked

Table 2 Sinclair's analysis of *budge*

'core' (always present)
  budge
'collocation' (strong association between two lexical items)
  refuse, could not, won't
'colligation' (strong association between the core and a grammatical category)
  negatives, modals
  'refusal' when the verb is intransitive
    e.g., *they won't budge*
  'inability' when the verb is transitive
    e.g., *two horses could not budge it*
'semantic prosody' (attitudinal, pragmatic meaning)
  expressing frustration at the refusal or inability
    e.g., *Do what they might, the British would not budge from their immigration policy*

Adapted from Sinclair, 2004: 142–147.

Table 3 Stubbs' analysis of *undergo*

'core':	*undergo*
'collocation':	*surgery, tests, treatment, change, training, test,* and so on
'colligation':	Preceded by passive or modal e.g., *forced to, must* Followed by adjective and abstract noun e.g., *further testing, major change*
'semantic preference':	Followed by nouns belonging to these sets: medical procedures; changes; nonmedical testing; other unpleasant things
'discourse prosody':	Indicates that a procedure is unpleasant Indicates that a procedure is involuntary

Adapted from Stubbs, 2001: 89–94.

is the notion of 'pattern grammar' (Francis, 1993; Francis *et al*., 1996, 1998; Hunston and Francis, 1999). This is a purely descriptive grammar of English that avoids abstract grammatical categories and that exploits the connection between phrase and meaning. Compiling a grammar of this kind involves the following:

- developing a notation that accounts for the colligational behavior of each lexical item to be included,
- collecting together all the lexical items that share a given notational sequence, and
- grouping those items according to meaning.

For example, the noun *fact* is frequently followed by a *that* clause that expands on the nature of the 'fact' concerned (that is, it is not simply a relative clause). A simple notation that captures this behavior is 'N that' (noun followed by *that* clause). Other nouns with the same notation include *assumption*, *idea*, *notion*, and *view*. These can be grouped according to meaning (Francis *et al*., 1998: 108–113):

- Nouns indicating speech or writing: *accusation, claim*, *declaration*, *explanation*, *insistence*, *message*, *objection*, *promise*, *recommendation*, *statement*, *warning*, and many others
- Nouns indicating ideas and thought processes: *assumption*, *belief*, *certainty*, *decision*, *faith*, *hypothesis*, *idea*, *knowledge*, *misunderstanding*, *presupposition*, *realization*, *speculation*, *theory*, *view*, *wish*, and many others
- Nouns indicating emotions: *amazement*, *concern*, *delight*, *fear*, *gratitude*, *happiness*, *indignation*, *joy*, *pleasure*, *regret*, *sadness*, *terror*, *worry*, and many others
- Nouns indicating evidence or signs that something is the case: *clue*, *demonstration*, *evidence*, *indication*, *proof*, *sign*, *symptom*, and others
- Nouns indicating likelihood: *chance*, *danger*, *guarantee*, *impossibility*, *likelihood*, *possibility*, *probability*, *risk*, and others
- Other nouns: *advantage*, *benefit*, *disadvantage*, *problem*, *consequence*, *result*, *reason*, and many others.

All words can be described in terms of the patterns they enter into, expressed as a series of elements. However, the link between pattern and meaning is most clearly made in relation to verbs, nouns, and adjectives. Another example is the pattern 'verb + noun phrase + *on* + noun phrase.' In traditional grammar, this would be described as a transitive verb followed by an adjunct. The pattern terminology makes it clear that the prepositional phrase beginning with *on* is directly related to the choice of verb. Examples of verbs with this pattern include these, from Francis *et al*. (1996: 403–410):

- Verbs meaning 'give something good': *bestow*, *confer*, *heap*, *lavish*, *press*, *settle*; e.g., *confer prestige on someone*
- Verbs meaning 'give something bad': *blame*, *dump*, *foist*, *impose*, *inflict*, *perpetrate*, *thrust*, and others; e.g., *blame a crime on someone*
- Verbs meaning 'talk about something': *advise*, *compliment*, *lecture*, *question*, *update*, and others; e.g., *advise someone on their actions*
- Verbs meaning 'put something somewhere': *cast*, *clip*, *load*, *place*, *put*, *sprinkle*, *throw*, and others; e.g., *sprinkle powder on something*
- Verbs meaning 'focus attention or effort': *center*, *concentrate*, *direct*, *fasten*, *fix*, *focus*, *pin*, *project*, *turn*, and others; e.g., *focus attention on someone*
- Verbs meaning 'building on something': *base*, *ground*, *predicate*, and others; e.g., *base a theory on facts*
- Verbs meaning 'spend money'; e.g., *blow*, *save*, *spend*, *waste*; e.g., *waste money on something*.

The outcome is a description of all the patterns in a given language, together with the lexical items that govern them.

Corpus Linguistics and Linguistic Theory

Corpus-Based Descriptions

As has been noted, corpus linguistics is essentially a methodology or set of methodologies, rather than a theory of language description. Essentially, corpus linguistics means this:

- looking at naturally occurring language;
- looking at relatively large amounts of such language;
- observing relative frequencies, either in raw form or mediated through statistical operations;
- observing patterns of association, either between a feature and a text type or between groups of words.

Reduced to its essence in this way, corpus linguistics appears to be 'theory neutral,' although the practice of doing corpus linguistics is never neutral, as each practitioner defines what is meant by a 'feature' and what frequencies should be observed, in line with a theoretical approach to what matters in language. Approaches to the use of a corpus that essentially rely on the existence of categories derived from noncorpus investigations of language are sometimes referred to as 'corpus based' (Tognini-Bonelli, 2001).

Studies of this kind can test hypotheses arising from grammatical descriptions based on intuition or on limited data. Experiments have been designed specifically to do this (Nelson *et al*., 2002: 257–283). For example, Meyer (2002: 7–8) describes work on ellipsis from a typological and psycholinguistic point of view that predicts that of the three possible clause locations of ellipsis in American spoken English, one will be much more frequent than the others. A corpus study reveals this to be an accurate prediction. On the other hand, the study of pseudo-titles mentioned in the section 'Languages and Varieties' shows how assumptions about language – in this instance about the influence of one variety of English on another – can be shown to be false. Biber *et al*. (1999: 7) comment that "corpus-based analysis of grammatical structure can uncover characteristics that were previously unsuspected." They mention as examples of this the surprisingly high frequency of complex relative clause constructions in conversation, and the frequency of simplified grammatical constructions in academic prose.

A clearer integration between linguistic theory and corpus linguistics is demonstrated by Matthiessen's work on probability (see the section 'Probability'). This work takes its categories from an existing description of English (Halliday's (1985) systemic-functional grammar), but the corpus study was more integral to the theory, as it was the only way that statements about probability of occurrence of each item in the system could be made with accuracy.

Corpus-Driven Descriptions

However, more radical challenges to language description can be found. Sinclair (1991, 2004) argues that the kind of patterning observable in a corpus (and nowhere else) necessitate descriptions of a markedly different kind from those commonly available. Both the descriptions and the theories that they in turn inspire are, in Tognini-Bonelli's (2001) terms, "corpus driven." Some of the challenges to tradition that corpus-driven theories involve are these:

- Lexis and grammar are not distinct, and grammar is not an abstract system underlying language
- Choice of any kind is heavily restricted by choice of lexis
- Meaning is not atomistic, residing in words, but prosodic, belonging to variable units of meaning and always located in texts.

Evidence for these claims is presented in the section 'Observing patterned behavior' above. The notion of pattern grammar focuses on the way that different lexical items behave differently in terms of how they are complemented. Grammatical generalizations about complementation cannot be made without describing that individual lexical behavior. Similarly, choice between features such as 'positive' and 'negative' depends to some extent on lexical item, as some verbs (such as *afford*) occur in the negative much more frequently than most. In other words, the probability of any grammatical category's occurring is strongly affected not only by the register but also by the lexis used. Finally, the evidence of phraseology is that it makes more sense to see meaning as belonging to phrases than to individual words. Findings such as these have led many writers to see a need for descriptions of language that are radically different from those currently available.

Sinclair (1991, 2004) proposes, for example, that meaning be seen as belonging to 'units of meaning,' each unit being describable in the way set out in Table 2. He criticized conventional grammar for distinguishing between structures (a series of 'slots') and lexis (the 'fillers'), such that it appears that any slot can be filled by any filler: there are no restrictions other than what the speaker wishes to say. This is clearly sometimes the case, and when it is, Sinclair

describes the language as following the 'open-choice principle.' Competing with that, however, is the 'idiom principle,' which is followed whenever units of meaning longer than a word are employed. In these cases, the speaker chooses the semantic prosody and the core word, and the rest occurs without choice.

Another theoretical approach to the nonrandomness of language in practice is Hoey's (2005) theory of 'lexical priming.' Like Sinclair, Hoey places lexis at the heart of his description of English, and like Sinclair he notes that words typically occur with specific collocations and in specific grammatical configurations. Hoey goes further in noting that certain words occur at the beginning of paragraphs or of texts or in lexical chains, with a greater (or lesser) frequency than would be predicted by random distribution, a phenomenon he calls 'textual colligation.' He argues that an individual's experience of words in context, over many years, 'primes' each lexical item for use in a particular collocational or colligational configuration or for playing a particular role in a text. This theory argues against absolutism in grammar and discourse: it cannot be said that a word or phrase belongs absolutely to a given class of items, but rather a word or phrase is primed to behave according to the norms of that category.

Hoey's work highlights an inevitable tension between the language experience of the individual, which is unique, and the desire to talk about language as a collective, shared experience, to describe 'a language' rather than 'an individual's language.' Teubert (2004a) raises similar issues in relation to corpus studies of meaning. Teubert sees the study of meaning as the main concern of corpus linguistics, and the area of linguistics in which the study of corpora is superior to other methods, because a corpus is a record of language as a social act (rather than as a psychological phenomenon), and meaning, too, is "a social phenomenon" (Teubert, 2004a: 97). On the other hand, meaning is not shared in an abstract sense: different individuals mean different things even when they use the same lexical item. Rather, Teubert argues, meaning is social because the meaning of a word or phrase resides in the sum of the ways in which that word or phrase has ever been used. He cites the word *school* (Teubert, 2004a: 190), which can be demonstrated to have been used in a wide variety of contexts, in each of which the word contributes something slightly different to the meaning of the discourse as a whole (e.g., *after school*, *tough time at school*, *school holidays*, *high school*, *medical school*, *skipped school*, *walk to school*). What we might think of as the 'dictionary' meaning of *school* is a generalization from all these instances, but each instance remains somewhat different from that generalization. To see what a word or phrase means, or what it meant at any time in history, we would have to examine each instance of the word. This is clearly impossible, but a corpus offers us a sample of instances, and the Internet, as a large if unplanned collection of texts, gives us even more.

In their different ways, these researchers all stress that descriptions of language – whether they try to account for rules or choices or meaning – are abstractions from a very large number of instances and that, as abstractions, they are always only partially correct. This interpretation arises from a focus on the lexical item as the primary object of study, which itself reverses the usual priorities found in linguistics.

Applications of Corpus Linguistics

Applied linguistics has been described as the use of knowledge about language to solve real-world problems. In recent years, the benefits of looking at large amounts of naturally occurring language, in the form of corpora, have been welcomed by applied linguists. Although corpora have many applications – notably forensic linguistics, stylistics, and critical discourse analysis – this section describes briefly just two of the more frequently encountered applications of corpus linguistics: language teaching and translation.

Language Teaching

Corpora have influenced language teaching in three distinct ways. Firstly, the findings from corpus research have been used extensively to improve reference materials for learners, such as dictionaries and grammars. Secondly, learners are increasingly being encouraged to explore corpora for themselves. Finally, corpus techniques have been applied to study of learners' language.

Since the publication in the mid-1980s of the first learners' dictionary based on corpus research (Sinclair *et al.*, 1987), corpora have become an indispensable resource for lexicographers and grammarians. Modern learners' dictionaries typically pay more attention to phraseology, and in particular to collocation, than previous ones did. Similarly, grammar books for learners pay more attention to register variation, to spoken usage, and to the role of lexis in grammar (Sinclair *et al.*, 1990; Biber *et al.*, 1999). To a lesser extent, course books have also changed, now placing more emphasis on collocation and phraseology than previously.

Corpora have influenced the method, as well as the content, of language teaching. Advanced learners are frequently invited to access corpora themselves and to engage in "data-driven learning" (Johns, 1991; Bernadini, 2000), in which they use a corpus to

make their own generalizations about language use. One of the consequences of this is that learners are exposed to all the complexity of a language, and the task of teaching explicitly every aspect of that language looks less viable than it did before. As a result, data-driven learning coincides happily with the view of language learning that stresses guided observation on the part of the learner rather than exposition on the part of the teacher (Willis, 2003; Bernardini, 2004).

Finally, the language of learners themselves has been studied extensively through the development of learner corpora (Granger, 1998), that is, corpora consisting of collections of written or spoken texts produced by learners of a language. These allow the learners' output to be compared with that of native speakers and for persistent errors in learner language to be identified. A common methodology is to identify features of language that occur significantly more or less frequently in the learner corpus than in a comparable corpus of native-speaker texts and to use such disparities as the starting point for more qualitative research. The features investigated include groups of words such as adverbials (Altenberg and Tapper, 1998) and modal auxiliaries (Aijmer, 2002), as well as more abstract categories, such as word class (Granger and Rayson, 1998) and sequences of part-of-speech tags (Aarts and Granger, 1998).

Translation

Corpora can be uses to train translators, used as a resource for practising translators, and used as a means of studying the process of translation and the kinds of choices that translators make. Parallel corpora are often used in these applications, and software exists that will 'align' two corpora such that the translation of each sentence in the original text is immediately identifiable. This allows one to observe how a given word has been translated in different contexts (see, for example, Teubert's work on *travail* and *work/labor* mentioned in the section 'Languages and Varieties'). One interesting finding is that apparently equivalent words – such as English *go* and Swedish *gå*, or English *with* and German *mit* (Viberg, 1996; Schmied and Fink, 2000) – occur as translations of each other in only a minority of instances. This suggests differences in the ways those languages use the items concerned.

More generally, examination of parallel corpora emphasizes that what translators translate is not the word but a larger unit (Teubert and Čermáková, 2004). Although a single word may have many equivalents when translated, a word in context may well have only one such equivalent. For example, although *travail* as an individual word is sometimes translated as *work* and sometimes as *labor*, the phrase *travaux préparatoires* is translated only as *preparatory work*. Thus, Teubert and Čermáková argue, *travaux préparatoires* and *preparatory work* may be considered to be equivalent translation units, whereas no such claim can be made for *travaux* and *work*.

As well as giving information about languages, corpus studies have also indicated that translated language is not the same as nontranslated language. Studies of corpora of translated texts have shown that they tend to have higher incidences of very frequent words and that they tend to be more explicit in terms of grammar (Baker, 1993). They may also be influenced by the structure of the source language, as was indicated in the discussion of *wh-* clefts in English and Swedish in the section 'Languages and Varieties.' In communities where people read a large number of translated texts, the foreign language, via its translations, may even influence the home language. Gellerstam (1996) notes that some words in Swedish have taken on the meanings of English that look similar and argues that this is because translators tend to translate the English word with the similar-looking Swedish word, thereby using the Swedish word with a new meaning, which then enters the language. One example is the Swedish word *dramatisk*, which used to indicate something relating to drama but which now, like the English word *dramatic*, also means 'substantial and surprising.'

Conclusion

Corpus linguistics is a relatively new discipline, and a fast-changing one. As computer resources, particularly web-based ones, develop, sophisticated corpus investigations come within the reach of the ordinary translator, language learner, or linguist. Our understanding of the ways that types of language might vary from one another, and our appreciation of the ways that words pattern in language, have been immeasurably improved by corpus studies. Even more significant, perhaps, is the development of new theories of language that take corpus research as their starting point.

See also: Corpora; Corpora of Spoken Discourse; Corpus Approaches to Idiom; Mark-up Languages: Text; Parsing and Grammar Description, Corpus-Based; Systemic Theory; Treebanks and Tagsets.

Bibliography

Aarts J & Granger S (1998). 'Tag sequences in learners corpora: a key to interlanguage grammar and discourse.' In Granger (ed.). 132–142.

Aijmer K (2002). 'Modality in advanced Swedish learners' written interlanguage.' In Granger S, Hung J & Petch-Tyson S (eds.) *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam: Benjamins. 55–76.

Aijmer K & Altenberg B (eds.) (1996). *Languages in contrast*. Lund: Lund University Press.

Allén S (1970). *Frequency dictionary of present-day Swedish, vol. 1: graphic words, homograph components*. Stockholm: Almqvist & Wiksell.

Altenberg B & Tapper M (1998). 'The use of adverbial connectors in advanced Swedish learners' written English.' In Granger (ed.). 80–93.

Baker M (1993). 'Corpus linguistics and translation studies: implications and applications.' In Baker, Francis & Tognini-Bonelli (eds.). 233–250.

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

Bernadini S (2000). 'Systematising serendipity: proposals for concordancing large corpora with language learners.' In Burnard L & McEnery T (eds.) *Rethinking language pedagogy from a corpus perspective*. Frankfurt: Peter Lang. 225–234.

Bernardini S (2004). 'Corpora in the classroom: an overview and some reflections on future developments.' In Sinclair J (ed.) *How to use corpora in language teaching*. Amsterdam: Benjamins. 15–38.

Biber D (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber D & Finegan E (2001). 'Diachronic relations among speech-based and written registers in English.' In Conrad & Biber (eds.). 66–83.

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

Carter R (2004). *Language and creativity: the art of common talk*. London: Routledge.

Carter R & McCarthy M (1995). 'Grammar and the spoken language.' *Applied Linguistics 16*, 141–158.

Conrad S & Biber D (2000). 'Adverbial marking of stance in speech and writing.' In Hunston S & Thompson G (eds.) *Evaluation in text: authorial stance and the construction of discourse*. Oxford: Oxford University Press. 56–73.

Conrad S & Biber D (2001). 'Multi-dimensional methodology and the dimensions of register variation in English.' In Conrad & Biber (eds.). 13–42.

Conrad S & Biber D (eds.) (2001). *Variation in English: multi-dimensional studies*. London: Longman.

Cowie A (ed.) (1998). *Phraseology: theory, analysis, and applications*. Oxford: Oxford University Press.

Francis G (1993). 'A corpus-driven approach to grammar: principles, methods and examples.' In Baker, Francis & Tognini-Bonelli (eds.). 137–156.

Francis G, Hunston S & Manning E (1996). *Collins Cobuild grammar patterns 1: verbs*. London: HarperCollins.

Francis G, Hunston S & Manning E (1998). *Collins Cobuild grammar patterns 2: nouns and adjectives*. London: HarperCollins.

Garside R, Leech G & McEnery A (eds.) (1997). *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Gellerstam M (1996). 'Translations as a source for cross-linguistic studies.' In Aijmer & Altenberg (eds.). 53–62.

Gledhill C (2000). *Collocations in science writing*. Tübingen: Gunter Narr Verlag.

Granger S (1998). 'The computer learner corpus: a versatile new source of data for SLA research.' In Granger (ed.). 3–18.

Granger S (ed.) (1998). *Learner English on computer*. London: Longman.

Granger S & Rayson P (1998). 'Automatic profiling of learner texts.' In Granger (ed.). 119–131.

Halliday M (1978). *Language as social semiotic: the social interpretation of language and meaning*. London: Arnold.

Halliday M (1985). *An introduction to functional grammar*. London: Arnold.

Halliday M & James Z (1993). 'A quantitative study of polarity and primary tense in the English finite clause.' In Sinclair J, Hoey M & Fox G (eds.) *Techniques of description: spoken and written discourse*. London: Routledge. 32–66.

Halliday M, Teubert W, Yallop C & Čermáková A (2004). *Lexicology and corpus linguistics*. London: Continuum.

Hoey M (2005). *Lexical priming: a new theory of words and language*. London: Routledge.

Hundt M (2004). 'The passival and the progressive passive: a case study of layering in the English aspect and voice systems.' In Lindquist H & Mair C (eds.) *Corpus approaches to grammaticalization in English*. Amsterdam: Benjamins. 79–120.

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

Hunston S & Francis G (1999). *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: Benjamins.

Hyland K (2000). *Disciplinary discourses: social interactions in academic writing*. London: Longman.

Hyland K (2001). 'Humble servants of the discipline? Self reference in research articles.' *English for Specific Purposes 20*, 207–226.

Johansson M (1996). 'Contrastive data as a resource in the study of English clefts.' In Aijmer K, Altenberg B & Johansson M (eds.) *Languages in contrast*. Lund: Lund University Press. 127–152.

Johns T (1991). '"Should you be persuaded": two samples of data-driven learning materials.' In Johns T & King P (eds.) *Classroom concordancing ELR journal 4*. University of Birmingham.

Leech G, Rayson P & Wilson A (2001). *Word frequencies in written and spoken English*. London: Longman.

Louw B (1993). 'Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies.' In Baker, Francis & Tognini-Bonelli (eds.). 157–176.

Mattheissen C (2005). 'Frequency profiles of some basic grammatical systems: an interim report.' In Thompson G & Hunston S (eds.) *System and corpus: exploring connections*. London: Equinox.

McCarthy M & Carter R (2002). 'Ten criteria for a spoken grammar.' In Hinkel E & Fotos S (eds.) *New perspectives in grammar teaching in second language classrooms*. Mahwah, NJ: Lawrence Erlbaum. 51–75.
Meyer C (2002). *English corpus linguistics: an introduction*. Cambridge: Cambridge University Press.
Nelson G, Wallis S & Aarts B (2002). *Exploring natural language: working with the British component of the International Corpus of English*. Amsterdam: Benjamins.
Schmied J & Fink B (2000). 'Corpus-based contrastive lexicology: the case of English with and its German translation equivalents.' In Botley S, McEnery A & Wilson A (eds.) *Multilingual corpora in teaching and research*. Amsterdam: Rodopi. 157–176.
Scott M (1996). *Wordsmith tools*. Oxford: Oxford University Press.
Scott M (2001). 'Mapping key words to problem and solution.' In Scott M & Thompson G (eds.) *Patterns of text: in honour of Michael Hoey*. Amsterdam: Benjamins. 109–128.
Semino E & Short M (2004). *Corpus stylistics: speech, writing and thought presentation in a corpus of English writing*. London: Routledge.
Sinclair J (1991). *Corpus concordance collocation*. Oxford: Oxford University Press.
Sinclair J (2003). *Reading concordances*. London: Longman.
Sinclair J (2004). *Trust the text: language, corpus and discourse*. London: Routledge.
Sinclair J *et al.* (1987). *Collins Cobuild English language dictionary*. London: HarperCollins.
Sinclair J *et al.* (1990). *Collins Cobuild English grammar*. London: HarperCollins.
Stubbs M (2001). *Words and phrases: corpus studies of lexical semantics*. Oxford: Blackwell.
Teubert W (2004a). 'Language and corpus linguistics.' In Halliday, Teubert, Yallop & Čermáková (eds.). 173–112.
Teubert W (2004b). 'When did we start feeling guilty?' In Weigand E (ed.) *Emotion in dialogic interaction: advances in the complex*. Amsterdam: Benjamins. 121–162.
Teubert W & Čermáková A (2004). 'Directions in corpus linguistics.' In Halliday, Teubert, Yallop & Čermáková (eds.). 113–166.
Tognini-Bonelli E (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.
Viberg A (1996). 'Cross-linguistic lexicology: the case of English go and Swedish gå.' In Aijmer & Altenberg (eds.). 153–184.
Willis D (2003). *Rules, patterns and words: grammar and lexis in English language teaching*. Cambridge: Cambridge University Press.

Corpus Studies: Second Language

R Reppen, Northern Arizona University, AZ, USA

Corpus linguistics has contributed to several areas of applied linguistics. In addition to core contributions in the areas of lexicography and grammar, corpus linguistics has also provided insights into the areas of register variation (e.g., spoken versus written language, across academic disciplines, stylistic variation), language change over time using historical or diachronic corpora, studies of gender differences, and, more recently, the area of second language studies (Reppen, 2001; Granger *et al.*, 2002; Granger, 2003). By using large, principled collections of naturally occurring language, corpus linguistics can accurately explore and describe linguistic characteristics and patterns associated with language use in different contexts (e.g., talking among friends, giving a formal speech, writing a friend, writing a research paper), across different speakers, and how language varies regionally. These descriptions can then be used to accurately describe patterns of variation and can also be used to inform pedagogy for second-language learners.

Corpora consist of large collections of spoken and/or written texts, are typically stored on computers, and are often grammatically annotated and/or marked up for certain text features (e.g., Biber *et al.*, 1998; Meyer, 2002; Reppen and Simpson, 2002). Because of their large size, often well over a million words, it is essential to have tools that allow users to effectively and efficiently search the corpora. There is a variety of computer programs available, ranging from concordancing software (e.g., MonoConc, WordSmith) that can generate word lists and identify specific words or combinations of words, to sophisticated programs that can perform comparisons that track features across a range of texts. Most users will interface with corpora through the use of concordancing software, most of which can be used with either an unannotated corpus or one that is annotated for grammatical or text features. A concordancing program allows users to generate word frequency lists, see target words in context, look for expressions, and also search for particular combinations, such as verb plus preposition, or what verbs frequently occur with complement clauses (if using a grammatically tagged corpus). Concordancing programs are useful