16 Computer-assisted Text and Corpus Analysis: Lexical Cohesion and Communicative Competence

MICHAEL STUBBS

0 Introduction

When we read or hear a piece of connected text, we may find the language used familiar or not, and correspondingly easy or difficult to follow. Difficulties in understanding a written or spoken text – such as a set of instructions, a textbook, a lecture, or a story in a conversation – can have many causes. However, by and large, we find a text easy to understand if it consists of familiar topics being talked about in familiar ways. If everything is totally familiar, of course, the text will strike us as boring or full of clichés. But there are limits to the rate at which we can take in new information, and we can understand connected text only if we are able to predict, at least partly, what is going to be said. Conversely, we find a text difficult to understand if it is lexically and semantically dense: that is, if there is too little repetition of vocabulary, if frequent topic changes mean that too much new vocabulary is being introduced too rapidly, and if too many of the words are unfamiliar or being used in unusual combinations.

These expectations of what is likely to be said – our knowledge of what is probable and conventional – can only come from other texts which we have read or heard in the past. This means that individual texts are interpreted against an intertextual background of norms of language use. These norms, which are expressed largely in recurring collocations of words, can be revealed by the computer-assisted analysis of large corpora. That is, we can compare what occurs in individual texts with what frequently occurs in large numbers of texts of different kinds.

In this chapter I will discuss methods for making such comparisons, under the following main topics:

- the contribution of words and phrases to text cohesion;
- the intertextual relations between texts;
- the extent to which our linguistic competence includes knowledge of norms of language use.

1 Data and Terminology

My main aim is to illustrate some computer-assisted methods of analyzing the use of words and phrases in texts and corpora, and for this, I require some simple terminology as follows.

A **text** is any stretch of naturally occurring language in use, spoken or written, which has been produced, independently of the analyst, for some real communicative purpose. A **corpus** is a large collection of computer-readable texts, of different text-types, which represent spoken and/or written usage. No corpus can be a fully representative sample of the whole language, but such collections can at least be designed to represent major dimensions of language variation, such as spoken and written, casual and formal, fiction and nonfiction, British and American, intended for different age groups, for experts and lay persons, and so on. **Large** means at least millions, and possibly hundreds of millions, of running words (tokens).

All examples of text fragments and phrases in this article are attested in such corpora. The frequency data are mainly from the Bank of English corpus created by COBUILD at the University of Birmingham in the UK. (COBUILD stands for Collins Birmingham University International Language Database.) This corpus has been used in the design of major dictionaries and grammars (including Cobuild 1995a; Francis et al. 1996, 1998). By the late 1990s, the corpus totalled some 330 million words, including fiction and nonfiction books, newspapers, and samples of spoken English. The corpus is available in different forms: I have here mainly used a 56-million word subcorpus which is available over the internet as CobuildDirect.[1] I have also used a database on CD-ROM (Cobuild 1995b), which was constructed from a 200-million word subcorpus. Sinclair (1991: 13–26) describes the early corpus development.

Other individual examples are from the LOB (Lancaster–Oslo–Bergen) and Lund corpora, and from the Longman–Lancaster corpus. For descriptions of these corpora, see Biber (1988: 66ff) and Summers (1993).[2]

Since I am going to compare the use of words and phrases in texts and corpora, I also need to make some terminological distinctions here. A **lemma** is a lexeme or dictionary headword, which is realized by a **word form**: e.g. the lemma TAKE (upper case) can be realized by the word forms *take*, *takes*, *took*, *taking*, and *taken* (lower-case italicized). Corpus work has shown that different forms of a lemma often have quite different collocational behaviour.

A **node** (the word form, lemma, or other pattern under investigation) co-occurs with *collocates* (word forms or lemmas) within a given *span* of word forms, for example 4:4 (four words to left and right). Position in the span can be given if relevant: e.g. N – 1 = one word to the left of the node, N + 3 = three words to the right. A **collocation** is a purely lexical and nondirectional relation: it is a node–collocate pair which occurs at least once in a corpus. Usually it is frequent co-occurrences which are of interest, and **typical collocates** of a node are given in diamond brackets, for word forms or lemmas, or for a set of semantically related words:

(1) untold <N + 1: damage, misery, . . . ; millions, riches, . . . >

(2) CAUSE <abstract nouns denoting "unpleasant things">

These examples are discussed in more detail below. Such sets are usually open-ended, and the relations probabilistic, but measures of typicality can filter out idiosyncratic collocates, and reveal the typical cases. (Statistical methods are discussed by Clear 1993; Stubbs 1995a; Barnbrook 1996.)

A **prosody** is a feature which extends over more than one unit in a linear string. Here I will refer to **discourse prosodies** which extend over a span of words, and which indicate the speaker's attitude to the topic. Unpleasant prosodies are more frequent, but pleasant prosodies do occur:

(3) BREAK out <"unpleasant things", such as: disagreements, riots, sweat, violence, war>

(4) PROVIDE <"valuable things", such as: aid, care, employment, facilities, food, housing, jobs, money, opportunities, security, services, support, training; an answer, data, information>

The concept of prosody in this sense was first proposed by Sinclair (1991: 74–5, 112). Louw (1993) provides the first detailed discussion, and Stubbs (1996) and Bublitz (1996) give other examples.

Finally, it has become fairly standard to distinguish between **cohesion** and **coherence** (Widdowson 1979: 146; Brown and Yule 1983: 24–5, 194–9). Cohesion refers to linguistic features (such as lexical repetition and anaphora) which are explicitly realized in the surface structure of the text: Halliday and Hasan (1976) provide a thorough account. Coherence refers to textual relations which are inferred, but which are not explicitly expressed. Examples include relations between speech acts (such as offer–acceptance or complaint–excuse), which may have to be inferred from context, or other sequences which are inferred from background nonlinguistic knowledge.

2 Lexical Cohesion: An Introductory Example

Here then is an initial example of the intertextual relations between a text fragment and typical language use, as documented in large corpora. It shows how a cohesive text is built up through the use of variations on typical collocations. As Sinclair (1991: 108) puts it: "By far the majority of text is made of the occurrence of common words in common patterns, or in slight variations of those common patterns." The text fragment is from a book on the environment published in 1990 in the UK:[3]

(5) Here the Green Party has launched its Euro-election campaign. Its manifesto, "Don't Let Your World Turn Grey", argues that the emergence of the Single European Market from 1992 will cause untold environmental damage. It derides the vision of Europe as "310 million shoppers in a supermarket". The Greens want a much greater degree of self-reliance, with "local goods for local needs". They say they would abandon the Chunnel, nuclear power stations, the Common Agricultural Policy and agrochemicals. The imagination boggles at the scale of the task they are setting themselves.

For many readers, the cohesion of this text fragment will be due both to repeated words and to familiar phrasings. It is sometimes thought that lexical cohesion is mainly due to chains of repeated and related words, such as:

(6) Green, Grey, Greens; Euro-, European, Europe; Party, election, campaign, manifesto; Market, shoppers, supermarket, goods

In an influential critique of attempts at text analysis, Morgan and Sellner (1980: 179–80) objected that such lexical chains are of no linguistic interest, but merely "an epiphenomenon of coherence of content." However, I will argue that lexical cohesion is not only a reflex of content, but that it is also due to the stringing together and overlapping of phrasal units.

In the text fragment, some of these units are simply fixed multiword phrases:

(7) the Green Party; the Single European Market; the Common Agricultural Policy; nuclear power stations

Other chunks are more complex to identify: they are variants on frequent combinations, such that certain words greatly increase the expectation that other words will occur. However, this assumes that we know the norms of co-occurrence in the language, and it is these norms that can be investigated only via the frequency of co-occurrences in large corpora. I will take a series of phrases, in the order in which they occur in the text, and show to what extent they are typical usages:

(8) **from (5):** has launched its Euro-election campaign

For example, the word form *launched* co-occurs with restricted sets of semantically related words. Native speakers might think initially of phrases such as *launched a satellite*, *lifeboats were launched*. However, the corpus data show that a much more frequent usage (about 50 times as frequent) is with abstract nouns, involving a plan, which may be military:

(9) launched <appeal, bid, campaign, programme, project, strategy; attack, offensive, invasion>

Most occurrences collocate with a time reference, especially a reference to a first, new, or recent launch, and/or (as here) a *has*-form which indicates present relevance of a recent event.

(10) **from (5):** cause untold environmental damage

Here, the corpus data show that the most frequent collocates of CAUSE (as a verb) are overwhelmingly unpleasant. I studied the collocates of CAUSE (verb and noun) in a 425,000-word corpus of texts about environmental issues (discussed by Gerbig 1996). Frequent collocates were:

(11) CAUSE <blindness, damage, danger, depletion, harm, loss, ozone, problems, radiation, warming>

If different corpora gave different results, then these unpleasant associations might be an artifact of the data, not a collocational property of the word. However, I also studied the 38,000 occurrences of CAUSE in a much larger corpus of 120 million words of general English (Stubbs 1995a). Amongst the 50 most frequent collocates within a span of 3:3, there were only words (most frequently abstract nouns) with unpleasant connotations. The most frequent were:

(12) CAUSE <problem(s) 1806, damage 1519, death(s) 1109, disease 591, concern 598, cancer 572, pain 514, trouble 471>

In addition, CAUSE often occurs in the syntactic structure verb + adjective + noun, such as:

(13) cause considerable damage; cause great problems; cause major disruption; cause severe pain; cause untold damage

The last example is the one in the text. In turn, *untold* is usually followed by an abstract noun denoting something bad and unpleasant, or a large number and/or a large amount of money:

(14) untold <damage, misery, problems, suffering; billions, riches>

A few cases are positive (*brought untold joy*): but in this context CAUSE is not used.

(15) **from (5):** a much greater degree of self-reliance

Other patterns are more variable again, but still detectable. In the corpus data, there were hundreds of examples of the pattern: *a* + quantity adjective + *degree of* + abstract noun. The most frequent adjectives were *greater* and *high*, as in *a far greater degree of clarity*, *a high degree of support*. After *greater*, almost all the nouns expressed positive ideas: e.g. *cooperation*, *democracy*, *success*.

(16) **from (5):** the imagination boggles at

Some words have very restricted uses: only *mind* and the semantically related *imagination* frequently co-occur with *boggles*:

(17) **from (5):** the scale of the task

The combination *the scale of the* is followed by abstract nouns (such as *challenge*, *operation*, *problem*) which refer back to a general discourse topic. Logically, the *scale* could be large or small, but *the scale of the* is almost always used of something very large, and usually something bad. Typical phrases are *underestimated the scale of the destruction* and *cannot cope with the scale of the fraud*.

(18) **from (5):** the task they are setting themselves

The task has no single anaphoric referent. *Task* is often used as a metalinguistic label to encapsulate a preceding stretch of text (see below on such vocabulary). Things one commonly sets oneself are goals which are challenging or demanding. Typical collocates are abstract nouns, such as *aim*, *challenge*, *goal*, *objective*, *standards*, *target*, or *task*.

(19) **from (17) and (18):** the scale of the task – the task they are setting themselves

Where chunks overlap with each other in this way, Hunston and Francis (1998: 68) talk of pattern flow.

We now have examples of the expectations generated by some of the individual words and phrases. A mechanism of text cohesion becomes clearer if we now look at these phrases from the text together, because we see that several have to do with the meaning "large size." There are explicit references to size in the text fragment (*310 million*, *greater*), but also implicit references. If a campaign is *launched*, the implication is that it is a major event. *Untold*, *boggles*, and *the scale of the* all usually co-occur with large numbers or large amounts. These patterns are not explicit in the text, but implicit in the intertextual references to norms of language use. Each individual pattern is probabilistic, but cumulatively the intertextual expectations convey "large size" as a discourse prosody distributed across the text.

3 Collocations and Cohesion

What follows? Collocational facts are linguistic, and cannot be explained away on grounds of content or logic. Such combinations are idiomatic, but not "idioms," because although they frequently occur, they are not entirely fixed, and/or they are semantically transparent. More accurately, such idiomatic combinations pose no problem for decoding, but they do pose a problem for encoding: speakers just have to know that expected combinations are *brought untold joy*, but *caused untold damage*. (Makkai 1972 and Fillmore et al. 1988: 504–5 draw this distinction.)

Much recent linguistics emphasizes creative aspects of language at the expense of predictable combinations, which nevertheless constitute a large percentage of normal language use. The pervasiveness of such conventionalized language use, the correspondingly large role played by memory, and the implications for fluent and idiomatic native speaker competence have, however, been emphasized by Bolinger (1976), Allerton (1984), Pawley and Syder (1983), Sinclair (1991), and Miller (1993).

Such observations concern probabilistic features of English. It is possible to have the "pleasant" combination *cause for celebration*, but vastly more frequent are combinations such as *cause for concern*. With the verb, there is nothing illogical (and nothing ungrammatical?) about the collocation ?*cause an improvement*, yet it seems not to occur. (What does occur is *make an improvement*, or *achieve*, *bring about*, *lead to*, *produce*, *result in*, and *secure an improvement*.) Such syntagmatic patterning is much more detailed than is generally shown in grammars: it stretches well beyond words and short phrases, and provides a relatively unexplored mechanism of text cohesion. However, as I have illustrated, such analysis cannot be restricted to isolated texts, since it requires an analysis of intertextual relations, and therefore comparison of the actual choices in a given text, typical occurrences in other texts from the same text-type, and norms of usage in the language in general.

The literature on cohesion tends to neglect the role of collocations. For example, Halliday and Hasan (1976), in the standard reference on cohesion in English, have only four pages on collocations and regard them as "the most problematical part of lexical cohesion" (1976: 284). However, the role of collocations in text cohesion is discussed by Kjellmer (1991) and Bublitz (1996, 1998). Moon (1994, 1998: 259) argues that semi-fixed phrases provide a way of presenting stereotyped ideas, which avoids explicit evaluation, but encodes shared schemas which are institutionalized in the culture. Sinclair (1996) provides further detailed examples of the kind of lexical, grammatical, and semantic relations which make such extended lexical units cohesive.

Conversely, the large literature on collocations and phrase-like units almost always regards them in their own right as linguistic units, and neglects their contribution to text cohesion. Early work on "word clustering" was done by Mandelbrot (who is nowadays more often associated with chaos theory), and as early as the 1970s he used a 1.6-million-word corpus to identify the strength of clustering between co-occurring words (Damerau and Mandelbrot 1973). More recent work (e.g. Choueka et al. 1983; Yang 1986; Smadja 1993; Justeson and Katz 1995) has used computer methods to identify recurrent phrasal units in natural text. Cowie (1994, 1999) provides useful reviews and discussions of principles.

These characteristics of language use – frequency, probability, and norms – can be studied only with quantitative methods and large corpora. However, cohesion (which is explicitly marked in the text) must be distinguished from coherence (which relies on background assumptions). Therefore, we also have to distinguish between frequency in a corpus and probability in a text. In the language as a whole, *launched an attack* is much more frequent than *launched a boat*. But if the text is about a rescue at sea, then we might expect *launched the lifeboat* (though *launched a plan* is not impossible). The probability of coming across a given word combination will be stable across the language: this is probability across a sequence of events. But this is not the same as the probability of a single event in a specific text: especially given that linguistic events are not independent of each other (unlike successive flips of a coin). Our linguistic competence tells us that one of these general semantic patterns (*launched* "a plan" or *launched* "a boat") is highly likely: but given what we know about the topic under discussion, we know which pattern is more likely in a given text.

4 Grammatical, Feasible, Appropriate, Performed

The significance of extended lexicosemantic units for a theory of idiomatic language use is discussed by Pawley and Syder (1983). They argue that native speakers know hundreds of thousands of such units, whose lexical content is wholly or partly fixed: familiar collocations with variants, which are conventional labels for culturally recognized concepts. Speakers have a strong preference for certain familiar combinations of lexis and syntax, which explains why nonnative speakers can speak perfectly grammatically but still sound nonnative.

A reference to Hymes's (1972) influential article on communicative competence can put this observation in a wider context. Hymes proposes a way of avoiding the oversimplified polarization made by Chomsky (1965) between competence and performance. Hymes not only discusses whether (1) a sentence is formally possible (= grammatical), but distinguishes further whether an utterance is (2) psycholinguistically feasible or (3) sociolinguistically appropriate. In addition, not all possibilities are actually realized, and Hymes proposes a further distinction between the possible and the actual: (4) what, in reality, with high probability, is said or written. In an update of the theory, Hymes (1992: 52) notes the contribution of routinized extended lexical units to the stability of text.

Whereas much (Chomskyan) linguistics has been concerned with what speakers *can* say, corpus linguistics is *also* concerned with what speakers *do* say. But note the *also*. It is misleading to see only frequency of actual occurrence. Frequency data become interesting when they can be interpreted as evidence for typicality, and speakers' communicative competence certainly includes tacit knowledge of behavioral norms.

5 Collocations and Background Assumptions

An important approach to discourse coherence has used the concept of semantic frames and schemas. For example, Brown and Yule (1983) discuss the background assumptions we make about the normality of the world: "a mass of below-conscious expectations" (1983: 62), which contribute to our understanding of coherent discourse. They argue that "we assume that" doors open, hair grows on heads, dogs bark, the sun shines. These assumptions depend in turn on expected collocations: in English, hair is blond, trees are felled, eggs are rotten (but milk is sour, and butter is rancid), we kick with our feet (but punch with our fists, and bite with our teeth), and so on. Many such examples go back to an early study of syntagmatic relations in German by Porzig (1934). Examples are often restricted to the small set of such items available to intuition, and their very banality contributes to our sense of a predictable and stable world. In an influential sociological discussion, Berger and Luckmann (1966) point to the importance of frequent "institutional formulae" in the construction of a taken-for-granted everyday reality.

However, it is important to distinguish between those collocations which are accessible to introspection and those which actually occur in running text. Both have to be studied, precisely because they reveal differences between intuition and behavior. For example, the very fact that KICK implies FOOT means that the words tend *not* to collocate in real text, since they have no need to. I checked over 3 million running words, and found almost 200 occurrences of KICK. But in a span of 10:10 (ten words to left and right), there were only half-a-dozen occurrences each of *foot* and *feet*, in cases where more precision was given:

(20) with his left *foot* he gave a wild *kick* against the seat

(21) she swam [. . .] with *kicks* of her thick webbed hind*feet*

Words often make general predictions about the content of surrounding text. Loftus and Palmer (1974) showed that words for "hit" trigger different assumptions, and affect perception and memory, when witnesses to a traffic accident are questioned in different ways about what they have seen, as in: *How fast were the cars travelling when they bumped* (versus *smashed*) *into each other*? Such assumptions do not arise from nowhere, but are created by recurrent collocations. In the 56-million-word corpus, I studied verbs in the semantic field of "hit." Collocates of HIT itself show its wide range of uses, often metaphorical and/or in fixed phrases (*hit for six*, *hit rock bottom*). In contrast, BUMP has connotations of clumsiness, and collocates such as *accidentally*, *lurched*, *stumbled*. COLLIDE is used predominantly with large vehicles, and has collocates such as *aircraft*, *lorry*, *ship*, *train*. SMASH has connotations of crime and violence, and has collocates such as *bottles*, *bullet*, *looted*, *police*, *windscreen*.

6 Collocations and Cultural Connotations

Such collocations contribute to textual coherence via the assumptions which they trigger. In a detailed study of such connotations, Baker and Freebody (1989) investigated the distribution of collocations in children's elementary reading books. They found that the adjective *little* was very frequent, and that 50 percent of the occurrences of *girl*, but only 30 percent of the occurrences of *boy*, collocated with *little* ($p < 0.01$). They argue (1989: 140, 147) that such frequent associations make some features of the world conceptually salient, but the associations are implicit, and appear to be a constant, shared, and natural feature of the world (cf. above on Berger and Luckmann 1966). Thus, *little* connotes cuteness, and its frequent collocation with *girl* conveys a sexist imbalance in such books. Such ideas ("girls are smaller and cuter than boys") are acquired implicitly along with the recurrent collocation.

Again, collocations can have such connotations only because patterns in a given text reflect intertextual patterns in the language. I studied 300,000 occurrences of the adjectives *little*, *small*, *big*, and *large*, and found that they occur in largely complementary distribution, with quite different uses and collocates (Stubbs 1995b). In particular, *little* has strong cultural connotations. The following facts are very simple, but not explicitly presented in any dictionary I have found. In the database constructed from a 200-million-word corpus (Cobuild 1995b), the most frequent noun to co-occur with *little* is GIRL, and the most frequent adjective to co-occur with *girl* is *little*. The phrase *little girl(s)* is nearly 20 times as frequent as *small girl(s)*, whereas *little boy(s)* is only twice as frequent as *small boy(s)*. *Little* typically occurs in phrases such as *charming little girl* (or *funny little man*), and *small* typically occurs in rather formal phrases such as *relatively small amount*.

What follows from such data? First, even on its own, one of the most frequent words in the language can convey cultural stereotypes, and this provides an intertextual explanation of why *little* has the connotations it does in phrases such as *Little Red Riding Hood*. In combination with other words, however, *little* conveys even stronger expectations. The combination *little old* is cute and folksy, or critical and patronizing; it can also be used purely pragmatically, with an atypical adjective–pronoun construction:

(22) this frail little old woman; the dear little old church; a ramshackle little old van; any weedy little old man

(23) little old New York; little old me

Of over 70 instances, selected at random from the corpus data, of *little old* before a noun, over half were in phrases such as *little old lady/ies* and *little old grandma*. The combination *little man* has two distinct uses. Both convey speaker attitude, one pejorative, and one approving:

(24) a ridiculous little man; an evil, nasty, frightful and revolting little man

(25) the little man against the system; little man versus Big Business; a victory for the little man

Second, paradigmatic oppositions (e.g. *little–big*, *old–young*) might appear to be permanently available in the language system. But coselection severely limits such choices in syntagmatic strings. There are stereotyped phrases such as *little old lady*, but combinations such as **little young lady* or ?*small old lady* are impossible or highly unlikely. Indeed it is frequent for paradigmatically contrasting items to co-occur (syntagmatically) within a text. Justeson and Katz (1991) discuss quantitative aspects of several adjective pairs including *large* and *small*, such as the tendency (highly statistically significant) of lexically antonymous adjectives to co-occur within a span of a few words, as in:

(26) from the *large* departmental store to the *small* shoe-mender

(27) a *large* area of the *small* kitchen

In summary: in terms of cohesion, the word *little*, especially in frequent collocations, allows a hearer/reader to make predictions about the surrounding text. In terms of communicative competence, all words, even the most frequent in the language, contract such collocational relations, and fluent language use means internalizing such phrases. In terms of cultural competence, culture is encoded not just in words which are obviously ideologically loaded, but also in combinations of very frequent words. (Cf. Fillmore 1992 on *home*.) One textual function of recurrent combinations is to imply that meanings are taken for granted and shared (Moon 1994).

7 Lexis and Text Structure

In this section, I will review some further aspects of lexical cohesion which I have not yet mentioned.

Some words function primarily to organize text: see Halliday and Hasan (1976) on general nouns which can refer to whole topics (such as *affair*, *business*, *matter*); Winter (1977) on cohesive lexical items (such as *conclude*, *fact*, *reason*, *subsequent*); Widdowson

(1983) on "procedural vocabulary"; and Tadros (1994) on "prospective vocabulary." These studies do not use computational techniques, though their lists can be used in such work. Yang (1986) identifies technical and subtechnical vocabulary by its distribution: technical words are frequent only in a restricted range of texts on a specialized topic, but not evenly distributed across academic texts in general; whereas subtechnical words (e.g. *accuracy*, *basis*, *decrease*, *effect*, *factor*, *result*) are both frequent and evenly distributed in academic texts, independent of their specialized topic. And Francis (1994) uses corpora to identify noun phrases typically used to encapsulate and evaluate topics in written argumentative texts (e.g. *this far-sighted recommendation*, *this thoughtless and stupid attitude*); such discourse labels often occur in frequent collocations, which may be recognizable as newspaper language (e.g. *reverse this trend*, *the move follows*, *denied the allegations*).

Words from given lexical fields will co-occur and recur in particular texts. For example, here are the ten most frequent content words (i.e. excluding very high-frequency grammatical words), in descending frequency, from two books:

(28) people, man, world, technology, economic, modern, development, life, human, countries

(29) women, women's, discrimination, rights, equal, pay, work, men, Act, government

Such lists fall intuitively into a few identifiable lexical fields, tell us roughly what the books are "about," and could be used as a crude type of content analysis. Work on the structural organization of vocabulary usually considers paradigmatic relations, but words in lexical fields can also be discovered by simple syntagmatic analysis. The classic work on lexical fields was done on German between the 1920s and 1940s by Trier (1931) and Weisgerber (1950): it is summarized by Ullmann (1957) and Lyons (1977).

Morris and Hirst (1991) identify topical units in texts via chains of word relations (such as synonymy, antonymy, part–whole) taken from a thesaurus. They implement, by hand, a procedure which can "delineate portions of text that have a strong unity of meaning," but claim that the procedure is computable (1991: 23, 29). Topic and content are signaled by vocabulary, which must provide at least some formal signals of text structure, since lexis is not distributed evenly across corpora or across individual texts. As Church and Mercer (1994: 10–11) put it, content words tend to appear in bunches, "like buses in New York City." If we divide a large corpus into 10,000-word segments, the occurrence of a given word, say *Kennedy*, will be distributed quite unevenly across the segments: perhaps several occurrences in two or three segments, but none at all elsewhere, and this uneven distribution is itself one mechanism of cohesion. Phillips (1985, 1989) therefore uses entirely automatic methods to study the distribution of lexis in the syntagmatic, linear stream of science textbooks. When we remember what a text is "about," we do not remember the syntactic structure: there are forms of organization to which grammatical classification is irrelevant. Phillips finds syntagmatic lexical sets, but, by carrying out the kind of objective, knowledge-free distributional analysis originally proposed by Harris (1952), he also finds that sets of words intercollocate. This shows distinct lexical networks within different chapters, and thus reveals semantic units not directly observable in the original text.

Even finer lexical clustering can be studied as follows. For the first few words of a text, all words occur for the first time. But very soon, words start to recur: that is, the number of word types (new words) rises more slowly than the number of word tokens (running words). Exceptions will occur only with texts of restricted kinds, such as a shopping list in which each word probably occurs just once. Such features of texts can be studied via their type–token ratio. On its own, this ratio provides only an overall characterization of a text. However, as new topics are introduced into a text, there will be bursts of new words, which will in turn start to recur after a short span. Youmans (1991, 1994) uses this observation to study the "vocabulary flow" within a text. He shows that if the type–token ratio is sampled in successive segments of texts (e.g. across a continuously moving span of 35 words), then the peaks and troughs in the ratio across these samples correspond to structural breaks in the text, which are identifiable on independent grounds. Therefore a markedly higher type–token ratio means a burst of new vocabulary, a new topic, and a structural boundary in the text. (Tuldava 1995: 131–48 also discusses how the "dynamics of vocabulary growth" correspond to different stages of a text.)

8 Observational Methods

This chapter has been mainly about empirical methods of studying lexis in texts and corpora. So I will end with some more general remarks on computer-assisted observational methods.

There are many aspects to the Saussurian paradox (Labov 1972: 185ff). In much recent linguistics, langue or competence is seen as systematic and as the only true object of study: but, since it is abstract ("a social fact" or an aspect of individual psychology), it is unobservable. Parole or performance is seen as unsystematic and idiosyncratic, and therefore, at best, of only peripheral interest: but, although concrete, it is observable only in passing fragments, and, as a whole, also unobservable. Mainstream linguistics – from Saussure to Chomsky – has defined itself with reference to dualisms, whose two poles are equally unobservable.

Observational problems arise also in connection with the traditional dichotomy between syntagmatic and paradigmatic. For Saussure (1916/1968: 171), the syntagmatic relation holds between items *in praesentia*, which co-occur in a linear string. A text is a fragment of parole, where instances of syntagmatic relations can be observed. However, we are interested in more than what happens to occur in such a fragment. A paradigmatic ("associative") relation is a potential relation between items *in absentia*, which have a psychological reality ("des termes *in absentia* dans une série mnémonique virtuelle", 1916/1968: 171). However, since paradigmatic relations are a virtual mental phenomenon, they are unobservable.

In an individual text, neither repeated syntagmatic relations, nor any paradigmatic relations at all, are observable. However, a concordance makes visible repeated events: frequent syntagmatic co-occurrences, and constraints on the paradigmatic choices. The co-occurrences are visible on the horizontal (syntagmatic) axis of the individual concordance lines. And the paradigmatic possibilities – what frequently recurs – are equally visible on the vertical axis: especially if the concordance lines are merely reordered alphabetically to left or right (Tognini-Bonelli 1996).

As a very brief illustration, here are examples of one of the patterns discussed above, in (15):

(30) a certain degree of humility
an enormous degree of intuition
a greater degree of social pleasure
a high degree of accuracy
a high degree of confidence
a large degree of personal charm
a mild degree of unsuitability
a reasonable degree of economic security
a reasonable degree of privacy
a substantial degree of association

This tiny fragment of data, extracted from a concordance, is not claimed in any way as representative: these are only ten examples from many hundreds. They simply illustrate that concordance lines make it easy to see that *degree of* is often preceded by a quantity adjective (the full concordance shows that by far the most frequent is *high*), and is often followed by an abstract noun (the majority of which express positive ideas). Concordances provide a powerful method of identifying the typical lexicogrammatical frames in which words occur.

The classic objection to performance data (Chomsky 1965: 3) is that they are affected by "memory limitations, distractions, shifts of attention and interest, and errors." However, it is inconceivable that typical collocations and repeated coselection of lexis and syntax could be the result of performance errors. Quantitative work with large corpora automatically excludes idiosyncratic instances, in favor of what is central and typical.

It is often said that a corpus is (mere) performance data, but this shorthand formulation disguises important points. A corpus is a sample of actual utterances. However, a corpus, designed to sample different text-types, is a sample not of one individual's performance, but of the language use of many speakers. In addition, a corpus is not itself the behavior, but a record of this behavior, and this distinction is crucial. Consider a meteorologist's record of changes in temperature. The temperatures are a sequence of physical states in the world, which cannot be directly studied for the patterns they display. But the record has been designed by human beings, so that it can be studied. The intentional design of the record can convert the physical states in the world into a form of public knowledge. (This example is from Popper 1994: 7.) And, developing Halliday's (1991, 1992) analogy, such temperature records can be used to study not only local variations in the weather (which are directly observable in a rough and ready way), but also longer-term variations in the climate, which are certainly not directly observable.

Chomskyan linguistics has emphasized creativity at the expense of routine, which is seen as habit and as the unacceptable face of behaviorism. Other linguists (such as Firth 1957 and Halliday 1992) and sociologists (such as Bourdieu 1991 and Giddens 1984) have emphasized the importance of routine in everyday life. Corpus linguistics provides new ways of studying linguistic routines: what is typical and expected in the utterance-by-utterance flow of spoken and written language in use.

Large corpora provide a way out of the Saussurian paradox, since millions of running words can be searched for patterns which cannot be observed by the naked eye (compare devices such as telescopes, microscopes, and X rays). We can now study patterns which are not visible directly to a human observer, but which are nevertheless stable across the language performance of many speakers. An elegant defense and detailed study of such patterns is provided by Burrows (1987: 2–3), who talks of:

> evidence to which the unassisted human mind could never gain consistent, conscious access. Computer-based concordances, supported by statistical analysis, now make it possible to enter hitherto inaccessible regions of the language [which] defy the most accurate memory and the finest powers of discrimination.

In this chapter, I have illustrated methods which can identify the intertextual patterns which contribute to the cohesion of individual texts. As Hymes (1972) argued thirty years ago, tacit knowledge of the probabilities of such patterns is a significant component of linguistic competence.

NOTES

For access to the Bank of English corpus, I am very grateful to colleagues at Cobuild in Birmingham. For permission to use other corpus materials, I am grateful to the Norwegian Computing Centre for the Humanities and Longmans Group UK Ltd. For comments on earlier drafts, I am grateful to Wolfram Bublitz, Andrea Gerbig, Gabi Keck, and Henry Widdowson. For corpus preparation and programming, I am grateful to Oliver Mason and Oliver Hardt.

1 CobuildDirect is available on-line, with access software, at http://titania.cobuild.collins.co.uk/form.html.

2 LOB consists of one million words of written British English; Lund consists of half a million words of spoken British English. These are very small corpora by modern standards, but carefully constructed, and still useful as reference corpora. The Longman–Lancaster corpus consists of 30 million words of written English, fiction and nonfiction. A useful further source is the 100-million-word British National Corpus, available on-line at http://thetis.bl.uk/lookup.html.

3 The book is *A Year in the Greenhouse* by John Elkington (London: Gollancz). A large part of the book is contained in the Longman–Lancaster corpus, reference LL 40433.

REFERENCES

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

Allerton, D. J. (1984). Three (or four) levels of word co-occurrence restriction. *Lingua*, 63: 17–40.

Armstrong, S. ed. (1994). *Using Large Corpora.* Cambridge, Mass.: MIT Press.

Baker, C. and Freebody, P. (1989). *Children's First School Books.* Oxford: Blackwell.

Baker, M., Francis, G., and Tognini-Bonelli, E. eds. (1993). *Text and Technology*. Amsterdam: Benjamins.

Barnbrook, G. (1996). *Language and Computers.* Edinburgh: Edinburgh University Press.

Berger, P. and Luckmann, T. (1966). *The Social Construction of Reality.* London: Allen Lane.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge: Cambridge University Press.

Bolinger, D. (1976). Meaning and memory. *Forum Linguisticum*, 1, 1: 1–14.

Bourdieu, P. (1991). *Language and Symbolic Power.* Oxford: Polity.

Brown, G. and Yule, G. (1983). *Discourse Analysis.* Cambridge: Cambridge University Press.

Bublitz, W. (1996). Semantic prosody and cohesive company: "somewhat predictable." *Leuvense Bijdragen: Tijdschrift voor Germaanse Filologie*, 85, 1/2: 1–32.

Bublitz, W. (1998). "I entirely dot dot dot": copying semantic features in collocations with up-scaling intensifiers. In R. Schulze, ed., *Making Meaningful Choices in English.* Tuebingen: Narr. 11–32.

Burrows, J. F. (1987). *Computation into Criticism.* Oxford: Clarendon.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press.

Choueka, Y., Klein, S. T., and Neuwitz, E. (1983). Automatic retrieval of frequent idiomatic and collocational expressions in a large corpus. *ALLC Journal*, 4, 1: 34–8.

Church, K. and Mercer, R. L. (1994). Introduction to the special issue on computational linguistics using large corpora. In Armstrong 1994: 1–24.

Clear, J. (1993). From Firth principles: computational tools for the study of collocation. In Baker et al. 1993: 271–92.

Cobuild. (1995a). *Collins Cobuild English Dictionary.* London: HarperCollins.

Cobuild. (1995b). *Collins Cobuild Collocations on CD-ROM*. London: HarperCollins.

Coulthard, R. M. ed. (1994). *Advances in Written Text Analysis*. London: Routledge.

Cowie, A. P. (1994). Phraseology. In R. E. Asher, ed., *The Encyclopedia of Language and Linguistics.* Oxford: Pergamon. 3168–71.

Cowie, A. P. (1999). Phraseology and corpora. *International Journal of Lexicography*, 12, 4: 307–23.

Damerau, F. J. and Mandelbrot, B. B. (1973). Tests of the degree of word clustering in samples of written English. *Linguistics*, 102: 58–75.

Fillmore, C. J. (1992). Corpus linguistics or computer-aided armchair linguistics. In Svartvik 1992: 35–60.

Fillmore, C. J., Kay, P., and O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of let alone. *Language*, 64, 3: 501–38.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–55. In *Studies in Linguistic Analysis*. Special Volume, Philological Society. Oxford: Blackwell. 1–32.

Francis, G. (1993). A corpus-driven approach to grammar: principles, methods and examples. In Baker et al. 1993: 137–56.

Francis, G. (1994). Labelling discourse: an aspect of nominal-group lexical cohesion. In Coulthard 1994: 83–101.

Francis, G., Hunston, S., and Manning, E. (1996). *Grammar Patterns 1: Verbs*. London: HarperCollins.

Francis, G., Hunston, S., and Manning, E. (1998). *Grammar Patterns 2: Nouns and Adjectives*. London: HarperCollins.

Gerbig, A. (1996). *Lexical and Grammatical Variation in a Corpus*. Frankfurt: Peter Lang.

Giddens, A. (1984). *The Constitution of Society*. Cambridge: Polity.

Halliday, M. A. K. (1991). Corpus studies and probabilistic grammar. In Aijmer and Altenberg 1991: 30–43.

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

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. London: Longman.

Harris, Z. (1952). Discourse analysis. *Language*, 28: 18–23.

Hoey, M. ed. (1993). *Data, Description, Discourse*. London: HarperCollins.

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

Hymes, D. (1972). On communicative competence. In J. Pride and J. Holmes, eds, *Sociolinguistics*. Harmondsworth: Penguin. 269–93.

Hymes, D. (1992). The concept of communicative competence revisited. In M. Pütz, ed., *Thirty Years of Linguistic Evolution*. Amsterdam: Benjamins. 31–58.

Justeson, J. S. and S. M. Katz (1991). Redefining antonymy: the textual structure of a semantic relation. In *Using Corpora*. Proceedings of 7th Annual Conference of the UW Centre for the New OED and Text Research. Oxford. 138–53.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology. *Natural Language Engineering*, 1, 1: 9–27.

Kjellmer, G. (1991). A mint of phrases. In Aijmer and Altenberg 1991: 111–27.

Labov, W. (1972). The study of language in its social context. In *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press. 183–259.

Loftus, E. F. and Palmer, J. C. (1974). Reconstruction of automobile destruction: an example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior*, 13: 585–9.

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

Lyons, J. (1977). *Semantics*. Cambridge: Cambridge University Press.

Makkai, A. (1972). *Idiom Structure in English*. The Hague: Mouton.

Miller, J. (1993). Spoken and written language. In R. J. Scholes, ed., *Literacy and Language Analysis*. Hillsdale, NJ: Erlbaum. 99–141.

Moon, R. (1994). The analysis of fixed expressions in text. In Coulthard 1994: 117–35.

Moon, R. (1998). *Fixed Expressions and Idioms in English*. Oxford: Clarendon.

Morgan, J. L. and Sellner, M. B. (1980). Discourse and linguistic theory. In R. J. Spiro, B. C. Bruce, and W. F. Brewer, eds, *Theoretical Issues in Reading Comprehension*. Hillsdale, NJ: Erlbaum. 165–200.

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17, 1: 21–48.

Pawley, A. and Syder, F. H. (1983). Two puzzles for linguistic theory. In J. C. Richards and R. W. Schmidt, eds, *Language and Communication*. London: Longman. 191–226.

Phillips, M. K. (1985). *Aspects of Text Structure*. Amsterdam: North-Holland.

Phillips, M. K. (1989). Lexical structure of text. *Discourse Analysis Monograph 12*. Birmingham: English Language Research.

Popper, K. R. (1994). *Knowledge and the Body–Mind Problem*. London: Routledge.

Porzig, W. (1934). Wesenhafte Bedeutungsbeziehungen. *Beiträge zur Geschichte der deutschen Sprache und Literatur*, 58: 70–97.

Saussure, F. de (1916/1968). *Cours de Linguistique Générale*. Paris: Payot.

Sinclair, J. McH. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. McH. (1996). The search for units of meaning. *Textus*, IX: 75–106.

Smadja, F. (1993). Retrieving collocations from text. Xtract. *Computational Linguistics*, 19, 1: 143–77.

Stubbs, M. (1995a). Collocations and semantic profiles. *Functions of Language*, 2, 1: 1–33.

Stubbs, M. (1995b). Collocations and cultural collocations of common words. *Linguistics and Education*, 7, 4: 379–90.

Stubbs, M. (1996). *Text and Corpus Analysis*. Oxford: Blackwell.

Summers, D. (1993). Longman/Lancaster English language corpus: criteria and design. *International Journal of Lexicography*, 6, 3: 181–95.

Svartvik, J. ed. (1992). *Directions in Corpus Linguistics*. Berlin: Mouton.

Tadros, A. (1994). Predictive categories in expository text. In Coulthard 1994: 69–82.

Tognini-Bonelli, E. (1996). Corpus Theory and Practice. Unpublished PhD dissertation. University of Birmingham, UK.

Trier, J. (1931). *Der deutsche Wortschatz im Sinnbezirk des Verstandes*. Heidelberg: Winter.

Tuldava, J. (1995). *Methods in Quantitative Linguistics*. Trier: WVT (Wissenschaftlicher Verlag Trier).

Ullmann, S. (1957). *The Principles of Semantics*. Oxford: Blackwell.

Weisgerber, L. (1950). *Vom Weltbild der deutschen Sprache*. Dusseldorf: Schwann.

Widdowson, H. G. (1979). *Explorations in Applied Linguistics*. Oxford: Oxford University Press.

Widdowson, H. G. (1983). *Learning Purpose and Language Use*. Oxford: Oxford University Press.

Winter, E. (1977). A clause-relational approach to English texts: a study of some predictive lexical items in written discourse. *Structional Science*, 6: 1–92.

Yang, H. (1986). A new technique for identifying scientific/technical terms and describing science texts. *Literary and Linguistic Computing*, 1, 2: 93–103.

Youmans, G. (1991). A new tool for discourse analysis: the vocabulary management profile. *Language*, 67, 4: 763–89.

Youmans, G. (1994). The vocabulary management profile: two stories by William Faulkner. *Empirical Studies of the Arts*, 12, 2: 113–30.