20 What can a corpus tell us about discourse?

Scott Thornbury

1. What is discourse?

The term *discourse* is both slippery and baggy: slippery because it eludes neat definition, and baggy because it embraces a wide range of linguistic and social phenomena. In this chapter the term will be limited to two basic senses – the formal sense: *discourse as connected text* (or *discourse*[1], for convenience) and the functional one: *discourse as language in use* (or *discourse*[2]). These two senses are well captured in the following definition:

> A piece of discourse is an instance of spoken or written language that has describable internal relationships of form and meaning … that relate coherently to an external communicative function or purpose and a given audience/interlocutor. Furthermore, the external function or purpose can only be properly determined if one takes into account the context and participants … in which the piece of discourse occurs.
>
> (Celce-Murcia and Olshtain 2000: 4)

The reference to the context and the participants can, of course, include not just the immediate *context of situation* but the larger social and cultural context as well. As Schiffrin (1994) puts it: 'To understand the language of discourse … we need to understand the world in which it resides' (1994: 419). Accordingly, I will invoke a third sense of the term discourse: *discourse as social practice* (Fairclough 1989), where the focus is 'not so much on how meanings are linguistically realized in texts, as on how they are socially constructed' (Widdowson 2007: xv). I will revisit this third sense of discourse (=*discourse*[3]*)* later in the chapter.

But the main focus of this chapter will be less on what corpus analysis can tell us about discourse as language in use (i.e. language in *context*), and more on what it can tell us about 'the describable internal relationships' of texts (i.e. language and its *co-text*). This is partly because the relationship between language and context is dealt with elsewhere in this volume (in particular chapters by Biber, on register, and Rühlemann, on pragmatics), and partly because – as we shall see in Section 3 – corpus linguistics is more comfortable handling *co-text* than it is *context*.

What, then, is subsumed under 'the describable internal relationships' of texts? For convenience, we can distinguish between two broad areas:

- *cohesion* across sentences and utterances, using grammatical and lexical devices, such as conjuncts, referring expressions and lexical repetition (Halliday and Hasan 1976);
- the *organisation* and *management* of discourse (both spoken and written), including the distribution of given and new information, topic management, the use of discourse markers, turn-taking, exchanges, scripts, rhetorical structures and macro-structures, such as narratives (Brown and Yule 1983; Coulthard 1985; McCarthy 1991).

2. What can a corpus tell us about discourse?

For reasons that will be discussed in Section 3, discourse analysts have been slower off the mark than, say, lexicographers or grammarians, in responding to the opportunities offered by corpora and the tools to investigate them with. Of course, few if any contemporary discourse studies are *not* informed by the analysis of collections of texts. But, as Partington (2004) notes, 'simply employing a corpus in one's research does not necessarily make it a study in Corpus Linguistics' (2004: 12). To qualify as such, analysts would need to use quantitative methods with the aim of producing findings that are both descriptive and explanatory.

The descriptive findings are generated by searching for particular discourse features in a corpus – typically a collection of texts of a specific register, but possibly a single extended text, such as a textbook or a novel – using computational means. Explaining the frequency, significance and use of these features generally involves reference to context, either the immediate co-textual environment, or to other texts or other corpora of texts. For, as Stubbs (2001a) reminds us, 'in corpus work, context means two rather different things: not only co-text (a short span of a few words within one single text), but also inter-text (repeated occurrences, often a very large number, of similar patterns across different, independent texts)' (2001a: 57). Thus, the analyst may compare and contrast an individual text, or a sub-corpus of texts of a specific type, with texts of another type, or with a larger and more general reference corpus. An example of this approach is Stubbs' own (1994) comparison of two school textbooks, using the Lancaster Oslo/Bergen corpus as a benchmark.

What, then, do corpora permit us to generalise about discourse, and, specifically, about the describable internal relationships in texts? Put another way, what is *quantifiable* about discourse?

In the same way that corpus-derived frequency information has revolutionised language description at the level of lexis and grammar, so too has the study of discourse hugely benefited from the kinds of quantitative data that corpora yield. Thus, researchers, such as Altenberg (1990), Stenström (1994) and Aijmer (1996), using the London–Lund Corpus of Spoken English, identified and quantified characteristic features of spoken interaction, including interactional signals, discourse markers and hedging devices. Gardner (1998) and O'Keeffe *et al.* (2007) have used spoken corpora to classify different kinds of response tokens. Biber *et al.* (1999), drawing on the British National Corpus, extend their description of English grammar to include a treatment of conversational structure, including the use of discourse markers and ellipsis, and thereby set out to 'explore the interface between grammar and discourse analysis, lexis, and pragmatics' (1999: 45).

In a similar spirit, the *Cambridge Grammar of English* (Carter and McCarthy 2006) includes a section called 'Grammar Across Turns and Sentences', in which 'the emphasis is on the grammar of texts and the part played by grammar in achieving textual coherence' (2006: 243). In this way, corpus studies have been instrumental in dissolving the rigid distinctions between grammar and discourse. 'Grammar becomes discourse when conventional sentence-based units of description fail to account for the facts' (McCarthy 1998: 82). Corpus analysis uncovers the facts.

Generalising from these facts, from a discourse perspective, typically involves identifying the micro-features of specific text types and from these extrapolating textual macro-features. For example, the distribution of tense, aspect and modality verb forms in academic papers, and the way that such forms correlate to specific textual functions, has been the subject of a number of studies (Crookes 1986; Swales and Najjar 1987; Swales 1990; Flowerdew 2002). Likewise, researchers have used small corpora to investigate the generic features of such registers as sports commentary (Ghadessy 1988), economics texts (Bondi 1999) and computer conferencing (Yates 1996). But, because the focus of such studies is mainly on the way that texts instantiate the contextual variables of specific discourse communities, these are more properly dealt with as *register* studies (see Biber, this volume).

Using corpus tools to identify what makes individual texts cohesive, on the other hand, or to track their internal organisation through the use of discourse markers, is more problematic. Corpus tools cannot easily detect cohesive ties, such as pronominal reference, unless they have been tagged as such. And even so, it is another matter to identify what a device is cohesive *with*. However, innovative procedures are being developed to overcome these problems. Biber *et al.* (1998) describe the use of 'an interactive text analysis program' that functions like a spellchecker: it searches tagged text corpora for targeted features, such as noun phrases, 'while retaining human decision-making for those difficult analyses that involve meaning distinctions' (1998: 113). Using this tool, the researchers were able to code and compare the characteristics of anaphoric referring expressions in different registers.

The identification of discourse markers (Schiffrin 1987) presents similar challenges. One approach is to search the corpus for pre-selected items. Pulcini and Furiassi (2004), for example, investigate the choice and distribution of discourse markers in a corpus of teacher–student interviews by starting with an inventory of such markers and then searching the corpus for occurrences. A more inductive approach involves specifying, not the items to be searched for, but the contexts in which to look. Tao (2003), for example, investigates turn-taking mechanisms by searching a corpus of transcribed conversation for the word or words immediately following the speaker-tag. By this means he is able to show that a relatively small repertoire of turn-initiators – such as *and*, *yeah*, *well*, *right* – are put to extensive use, for both cohesive and pragmatic purposes.

Biber, Conrad and Cortes (2004) search corpora of two different registers – classroom teaching and textbooks – for frequently occurring word sequences ('lexical bundles') and find that many, such as *going to talk about* and *has to do with the*, have a discourse organising function. Their relative frequency in classroom teaching leads the researchers to conclude that 'these lexical bundles serve as discourse framing devices: they provide a kind of frame expressing stance, discourse organization, or referential status, associated with a slot for the expression of new information relative to that frame' (2004: 400). Again, such a conclusion depends not on frequency data alone, but on the use of concordance data in order both to infer the functional categories of these bundles, and to explain their relative frequency in different registers.

More amenable to corpus analysis are features of lexical cohesion (Halliday and Hasan 1976) including *reiteration* (the direct and indirect repetition of words, the use of synonyms, near synonyms and general terms) and *collocation*. With regard to collocation, corpus tools can, of course, easily identify words in a corpus that co-occur (and co-occur with more than chance frequency): this is one of their more robust functions. But they are limited to a relatively short co-textual span: typically from one to five words either side of the node. They are less sensitive to co-occurrences over larger stretches of text, even in adjacent sentences. (Some corpus tools, such as *Sketch Engine*, developed by Adam Kilgarriff, Lexical Computing Ltd, do however allow collocation searches over much wider spans, and developments in semantic tagging may soon allow researchers to capture more extensive meaning networks encoded in the lexis; see Walter, this volume, for an illustration of a *Sketch Engine* search.)

Corpus tools are better suited to identifying and tallying instances of reiteration, both direct and indirect, in a text. Frequency lists, after all, are simply a record of a text's – or a corpus of texts' – lexical repetitions. From the point of view of topical cohesion, a list of the words that are *key* in the texts may be more revealing than a simple frequency list. *Keyness* is defined as 'a quality words may have in a given text or set of texts, suggesting that they are important, [that] they reflect what the text is really about' (Scott and Tribble 2006: 73). Keyword analysis allows the analyst to explore 'not just how sentences are structured but how whole sections of text flow and move' (2006: 7).

Stubbs (2001a) shows how this textual 'flow' is achieved through the recurrence not just of individual words or their derivatives but of 'lexico-semantic units', including collocations and other formulaic lexical combinations, thereby creating 'a relatively unexplored mechanism of text cohesion' (2001a: 120). Similarly, Hoey (2005), building on his previous studies (e.g. 1991) of how chains of related lexis form *textual collocations* that ripple through whole texts and thereby create coherence, uses corpus data to identify which words frequently occur in such chains, and are thus *primed* for textual cohesion. Of particular relevance to discourse analysis is his claim that 'every lexical item (or combination of lexical items) is capable of being primed (positively or negatively) to occur at the beginning or end of an independently recognised "chunk" of text' (2005: 129), an effect that he terms *textual colligation*.

Hoey stresses that such primings are genre-specific. This point is demonstrated in a study of business communication (Scott and Tribble 2006) where a single keyword (*hope*) was found to correlate with the same discourse moves and to occur in similar textual environments across a number of texts in a small corpus of business correspondence. The researchers conclude that 'a combination of KW [keyword] analysis and discourse analysis offers teachers and students a powerful way of coming to an understanding of how language is used in professional settings' (2006: 109).

To similar ends Biber, Csomay *et al.* (2004) use a more elaborated approach in order to identify what they call Vocabulary-Based Discourse Units (VBDUs), 'based on the assumption that different discourse units tend to use different sets of words, reflecting shifts in topic and purpose' (2004: 24). Using a computer program that progressively scans sequences of text for evidence of lexical repetition, VBDUs were identified in a variety of genres. These units were then subjected to closer scrutiny, in order to identify any distinctive linguistic properties, and to map these on to their situational, social and cognitive functions. Using cluster analysis, seven VBDU types were identified, such as *extreme oral narrative* and *literate + content-focused*, with a view to 'studying how sequences of DU-types work together in different registers, supporting different major rhetorical patterns' (2004: 38).

Flowerdew (2003) also combines genre analysis and corpus linguistics – specifically, a keyword analysis – in order to investigate the problem-solution pattern in two corpora of technical academic writing, one by professionals and the other by students. Her particular focus is on the language of *appraisal* (Martin and Rose 2003). Stubbs (2007) approaches the same goal – of mapping specific semantic units on to the discourse structure of particular text-types – from the perspective of phraseology. His claim is that 'many frequent phrasal constructions have textual functions of evaluation and information management' (2007: 182), and goes on to argue that 'studies of phraseology must combine corpus analysis (which phrasal constructions are frequent in the corpus?) and textual analysis (how do these constructions organize individual texts?)' (2007: 182).

The above studies are concerned mainly with using corpus tools to identify and describe the internal relationships in texts (what, earlier, we termed *discourse*^{1}*)*. Corpus-based studies that focus on *discourse as language in use* (or *discourse*^{2}) and which 'take[.] into account the context and participants … in which the piece of discourse occurs' (Celce-Murcia and Olshtain 2000: 4) require the kind of contextual data that few large corpora provide (see the next section). Nevertheless, studies combining corpus data and other research tools, such as discourse completion tests (DCTs), provide insights into how the corpus data are realised in specific contexts and between specific participants (see, for example, Schauer and Adolphs 2006). Small corpora, gathered in restricted contexts, allow closer tracking of contextual factors. McCarthy (2000), for example, used two contextually coded extracts from the much larger CANCODE spoken corpus to investigate small talk at the hairdressers' and during driving lessons. Kuiper and Flindall (2000) gathered a small corpus of exchanges at supermarket checkouts in order both to describe and to explain the structure of these exchanges. Studies like these, that combine the thick contextual descriptions of ethnography with the quantifiable linguistic data of a corpus, offer useful methodological models for investigating discourse in action.

What is notable about all these studies is the way that they combine corpus-based procedures with research methods from other disciplines, such as genre analysis, phraseology, pragmatics and ethnography. Research into the third level of discourse, *discourse-as-social-practice*, is similarly eclectic in its methods. Stubbs (1996), for example, situates his corpus research firmly within the critical discourse analysis paradigm. In one study (1996), he uses quantitative methods to analyse patterns of transitivity in two school textbooks, in order to reveal their ideological biases. Elsewhere, however, he leans more in the direction of stylistics: in Stubbs (2005), using corpus tools again, he peels back the thematic sub-text of Joseph Conrad's *Heart of Darkness.* The value that a corpus-based methodology brings to these studies is that it 'helps to ensure that analysts do not merely pick evidence to fit their preconceptions' (Stubbs 1996: 154). This does not mean, though, that *discourse*^{3}-analysis is purely objective. As Teubert (2007) makes clear: 'The generation of relevant concordances and the statistical analysis of this data is never more than a first step. What the evidence means is a matter of interpretation. Without interpretation any study in corpus linguistics would be incomplete' (2007: 124–5). Teubert's own study (2000) of the language of British Euroscepticism, as expressed on websites opposed to the European Union, is a case in point. So, too, is the study by Koteyko *et al.* (2008) on media and government discourses about the 'superbug' (a particularly virulent micro-organism found in hospitals), where the value of using concordance programs to search corpora was that 'this allowed us to validate qualitative "hunches" using quantitative data' (2008: 239). Likewise, a study by Baker (2006), again using concordance data, of how newspaper texts construe refugees, leads him to conclude that

> the patterns of language which are found (or overlooked) [in a corpus] may be subject to the researcher's own ideological stance. And the way that they are interpreted may also be filtered through the researcher's subject position. This is true of many other, if not all, forms of discourse analysis. However, the corpus-based approach at least helps to counter some of this bias, by providing quantitative evidence of patterns that may be more difficult to ignore.
>
> (Baker 2006: 92)

3. What are the limitations of using a corpus in the study of discourse and how might we overcome them?

If discourse analysts have been slow to embrace the opportunities offered by corpus linguistics, this is in part due to the perception that corpora consist mainly of de-contextualised text fragments, assembled from a fairly random range of sources – an inevitable consequence of their original, primarily lexicographical, purpose. Discourse analysis, on the other hand, requires whole texts, often of the same type, as its database. And the bias, in many early corpora, towards written texts meant they were of little practical use to researchers of spoken language.

Now, however, specialised corpora of specific registers, including spoken language, have proliferated, and most corpora of general English, including many that are freely available online, are tagged for text-type and register. Most also allow access to more of each text than simply concordances of individual words. Nevertheless, analysts still lack tools that will perform many of the kinds of operations that are traditionally done manually. As Biber *et al.* (1998) point out, many features of connected text – such as the distribution of given versus new information, or the identification of pronoun referents and of other cohesive devices – cannot be detected automatically.

But there is a more fundamental problem facing the discourse analyst. While corpus tools allow researchers to track, tally and plot the surface features of discourse – such as its linking devices, discourse markers and instances of lexical repetition – these remain simply that: surface features. They do not necessarily correlate with, or explain, the underlying semantic relations between parts of a text, including those which account for the text's coherence and its generic structure. This is a limitation of the study of cohesion in general, and one that Halliday and Hasan (1976) were well aware of:

> Cohesion expresses the continuity that exists between one part of the text and another. It is important to stress that continuity is not the whole of texture. The organization of each segment of a discourse in terms of its information structure, thematic patterns and the like is also part of its texture … no less important than the continuity of one segment to another.
>
> (Halliday and Hasan 1976: 299)

Where such features leave lexical traces (as, for example, in the form of lexical chains), a corpus-based analysis is warranted (as Hoey 1991 and Biber, Csomay *et al.* 2004 have demonstrated), but such an analysis can only complement and not replace a more interpretative approach. In the end, discourse is more than words. As Baker (2006) warns, 'at present, a great deal of corpus-based discourse analysis is still focused at the lexical level. The challenge to future researchers is to find way to make grammar- and semantic-based analysis of corpora a more feasible option' (2006: 174).

Even were such means available, they would be of little practical use in the absence of detailed information about *context*. The lack of thick data relating to the contextual conditions of text production and interpretation continues to handicap the application of corpus analysis tools to the analysis of discourse. As Baker (2006) observes:

> Questions involving production such as who authored a text, under what circumstances, for what motives and for whom, in addition to questions surrounding the interpretation of a text: who bought, read, accessed, used the text, what were their responses, etc. can not be simply answered by traditional corpus-based techniques.
>
> (Baker 2006: 18)

Baker's concerns reflect Widdowson's (2004) well-rehearsed complaint that 'corpus analysis does not … account for context' (2004: 124), and that corpus linguists 'cannot … directly infer contextual factors from co-textual ones, and use textual data as conclusive evidence of discourse' (2004: 126). While this argument is hotly contested (see, for example, Stubbs 2001b), the use of smaller, more localised corpora (as was mentioned in the last section), where contextual information is rigorously specified, is an attempt to deal with the problem of lack of context.

In the end, quantitative data alone are not going to answer all – or even, any – of the questions that analysts bring to the study of discourse. As we have seen, a combination of computation and interpretation, in mutually informing cycles of investigation, and drawing on a variety of related disciplines – offers the most promising way forward. The next section demonstrates how this can work in practice.

4. How does a corpus-based approach work in practice?

It is now time to demonstrate how the application of corpus linguistics to discourse analysis might work in practice, with a view not only to validating the procedures involved, but to confronting some of the problems outlined in the previous section. The approach is essentially a bottom-up and inductive one: starting with frequency lists and word searches, the analyst identifies regularities in a corpus of texts of the same provenance and register, with a particular focus on those features that offer evidence of the internal relationships of the text-type in question. Studying these data, the researcher then constructs a provisional schematic for the overall structure of the text-type, which can then be checked against individual instances, and refined if necessary. Using the resulting description, and taking into account the contextual and cultural factors in which the texts are produced and interpreted, the researcher is then in a position to speculate as to how the formal features of the texts encode their communicative and social functions.

With these ends in mind, a small corpus (10,000 words) of teenage written narratives, hereafter referred to as the Cringe Text Corpus, was compiled, using an online teenage magazine as the source. The corpus consists of 143 short narratives of this type:

> One day I was walking in the park, and I saw a major babe and wanted to impress him. I started running in the sand volleyball courts and ran straight into the net! I fell flat on my back and started crying. He started laughing at me and it was terrible!

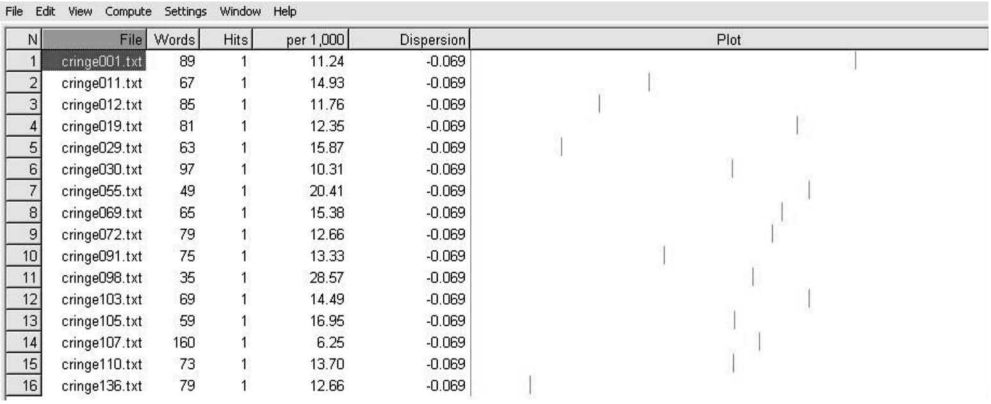
With the aim of identifying the internal relationships, including the generic discourse structure of the texts that comprise this corpus, *WordSmith Tools* (Scott 2008) was used, first to compile a list of the most frequent words in the corpus, and then to search these for linkers, specifically coordinating and subordinating conjunctions, and linking adjuncts. Because the linkers in the corpus were not tagged as such, a concordance program was then used in order to eliminate instances of linkage at the phrase level, such as *my best friend and I*, or polysemes, such as *I was so embarrassed*; *I ate too much*). Table 20.1 shows the linkers that occurred at least four times or more in the Cringe Text Corpus, including the number of instances in which the linker occurred at the beginning of a sentence.

The items in Table 20.1 are all single lexemes: a more thorough search, using a concordancer, was needed in order to identify combinations of linkers, such as *and then* (n = 6), as well as phrasal and clausal adjuncts, such as *all of a/the sudden* (n = 10) or combinations with *worse,* such as *even worse* and *to make matters* worse (n = 7). A word search also established that the following linkers (taken from lists in the *COBUILD English Grammar*; Sinclair 1990) do *not* occur in the Cringe Text Corpus: *furthermore*, *moreover*, *however*, *yet*, *nevertheless*, *therefore*, *hence*, *thus*, *consequently*, *secondly* or *thirdly*. (On the other hand, all these items did appear at least once in a corpus of academic journal abstracts.) The prominence of so many coordinating conjunctions, especially *and*, in the Cringe Text Corpus, suggests a markedly paratactic syntax (where sequenced clauses have equal status), in contrast to the more hypotactic style of the academic abstracts (where subordinate clauses are frequent). With regard to position in the sentence, only one linking adjunct (*then*) showed a slight preference for the sentence-initial slot.

As well as showing an item's position in a sentence, corpus tools can also display, graphically, where individual items occur in relation to the whole text. This is done through the use of a *dispersion plot*. Figure 20.1, for example, shows the distribution of all the instances of *sudden** (i.e. *suddenly*, *all of a sudden*) in the corpus, where the position of

Table 20.1 Linkers with a frequency of $\geq$ 4 in the Cringe Text Corpus

N	*Word*	*Freq.*	*%*	*Sentence initial*	*Texts (n = 143)*
1	and	366	3.37	1	136
2	so	73	0.67	6	63
3	but	53	0.49	4	45
4	then	26	0.25	15	25
5	finally	10	0.09	3	9
6	later	6	0.06	2	6
7	suddenly	6	0.06	1	6
8	though	4	0.04	0	4
9	too	4	0.04	0	4



N	File	Words	Hits	per 1,000	Dispersion	Plot
1	cringe001.txt	89	1	11.24	-0.069	
2	cringe011.txt	67	1	14.93	-0.069	
3	cringe012.txt	85	1	11.76	-0.069	
4	cringe019.txt	81	1	12.35	-0.069	
5	cringe029.txt	63	1	15.87	-0.069	
6	cringe030.txt	97	1	10.31	-0.069	
7	cringe055.txt	49	1	20.41	-0.069	
8	cringe069.txt	65	1	15.38	-0.069	
9	cringe072.txt	79	1	12.66	-0.069	
10	cringe091.txt	75	1	13.33	-0.069	
11	cringe098.txt	35	1	28.57	-0.069	
12	cringe103.txt	69	1	14.49	-0.069	
13	cringe105.txt	59	1	16.95	-0.069	
14	cringe107.txt	160	1	6.25	-0.069	
15	cringe110.txt	73	1	13.70	-0.069	
16	cringe136.txt	79	1	12.66	-0.069	

Figure 20.1 Dispersion plot for *sudden** in the Cringe Text Corpus.

the item in each of the sixteen texts in which it occurs is visually represented by a short vertical line in the column headed *Plot*.

Since the plot has been standardised so that each text file appears to be of the same length, it is possible to compare the point of occurrence of the search item across multiple texts. It is evident, for example, that the majority of instances of *sudden** (ten out of sixteen) occur at, or just after, the midway point in the texts. Such information provides a clue as to the discourse structure of the texts in this particular corpus – the fuller implications of which will be explored shortly.

We have already noted that, without a great deal of manual tagging, the cohesive properties of referring expressions, such as pronouns and demonstrative determiners, are not easily tracked using corpus tools. However, a search of the occurrences of *this* in the Cringe Text Corpus did identify a usage that was sufficiently frequent to qualify as a potential generic feature. Of the twenty-seven instances of *this* in the corpus, twenty are non-referring. That is to say, the noun phrases that they premodify are first mentions in the text. For example:

```
I was flirting with this guy online
One day I told this girl which guy I had a crush on,
and there was this really hot guy taking our money
I wore this skirt that was long and flared out.
```

Of these twenty occurrences, eight follow the pattern '*this [really/totally][hot/cute]* + male' as in *this really hot guy*. Of the remaining seven examples of *this* + NP in the corpus, five form part of the cluster *to this day,* leaving only two that have anaphoric reference. We can conclude that, on this evidence, and to use Hoey's (2005) terminology, *this* is not *primed* for endophoric reference (either anaphoric or cataphoric) in this kind of text.

So far we have been looking at instances of grammatical cohesion – principally conjunction and demonstrative reference. Now we turn our attention to lexical cohesion. As a preliminary stage, a keyword analysis (again using *WordSmith Tools*) was performed, in order to identify words that were unusually frequent in the entire corpus (for detailed coverage of keywords, see Scott, this volume). The top thirty keywords in the Cringe

Text Corpus, excluding function words, as measured against the BNC (world edition, 2001) are displayed in Table 20.2.

These thirty words alone are a strong indicator as to the thematic content of the narratives that comprise the Cringe Text Corpus. The fact that many are semantically related, either because they belong to the same lexical set (*school*, *locker*; *walked*, *ran*) or to the same word family (*friend*, *friends*, *friend's*; *walking*, *walked*), or because they are synonyms (*mortified*, *embarrassed*; *boyfriend*, *crush* (in this context, an informal term for *boyfriend*); *fell*, *tripped*) or because they are collocates (*cute* + *guy*; *fell* + *butt*), is an indicator that the texts that constitute the corpus have what Hasan (1989) terms *texture*: 'The texture of a text is manifested by certain kinds of semantic relations between its individual messages' (1989: 71). The semantic relations are typically instantiated in the form of cohesive *chains*, of which Hasan describes two types. An *identity chain* is a set of items that are co-referential: every member of the set refers to the same person or event. Hasan notes that, in short narratives, identity chains typically run the length of the whole text. In the Cringe Text Corpus narratives, identity chains are mostly realised in the form of the first-person narrator: the pronouns *I* and *my* are the first and third most frequent words in the corpus, together comprising over 11 per cent of all word tokens, and occurring in all but one of the 143 texts.

The items in a *similarity chain*, on the other hand, 'belong to the same general field of meaning, referring to (related/similar) actions, events, and objects and their attributes' (Hasan 1989: 85). A possible similarity chain in the Cringe Text Corpus might be *tripped – fell – butt*. It would need a more fine-grained search to confirm whether this is, in fact, the case. So, while a list of keywords is not in itself a semantic network, it provides the raw data out of which such a network might be constructed.

A short (typically two to six) chain of words that are related simply because they commonly co-occur is called a *cluster* (Scott 1997), also known as *lexical bundles* (Biber *et al.* 1999) or *n-grams* (Fletcher 2003/8) (see Greaves and Warren, this volume). A cluster analysis complements a keyword search, especially at their points of intersection, as it identifies typical contexts in which certain keywords recur. Thus, the four-word clusters that appear five times or more in the Cringe Text Corpus suggest that certain patterns

Table 20.2 Key content words in the Cringe Text Corpus

N	*Word*	*Freq.*	*Keyness*	*No. of texts (n = 143)*	*N*	*Word*	*Freq.*	*Keyness*	*No. of texts (n = 143)*
1	crush	71	872.39	49	16	embarrassed	18	136.06	18
2	mom	25	290.07	16	17	guys	17	131.19	12
3	mortified	20	270.29	20	18	boyfriend	16	127.76	12
4	friends	53	269.29	43	19	locker	12	122.34	9
5	friend	52	255.95	38	20	fell	27	120.50	26
6	laughing	30	223.73	24	21	butt	13	120.41	12
7	guy	30	203.24	23	22	everyone	28	118.93	27
8	really	63	202.28	49	23	cute	12	117.72	12
9	day	66	188.80	54	24	talking	26	102.59	21
10	bathroom	23	161.85	16	25	went	41	100.82	33
11	started	38	158.42	31	26	tripped	10	96.04	9
12	school	49	157.35	39	27	saw	31	94.02	29
13	walking	27	145.37	23	28	ran	19	79.98	19
14	friend's	17	142.99	13	29	walked	19	78.09	18
15	hot	29	142.09	25	30	brother	18	76.72	13

Note: the keyness is calculated according to log likelihood.

(and themes) recur regularly enough to be generic, and that they are talked about in the same way: the writers experience embarrassment when unexpected events, including falls, occur and which are witnessed by their boyfriend (*crush*). (See Table 20.3.)

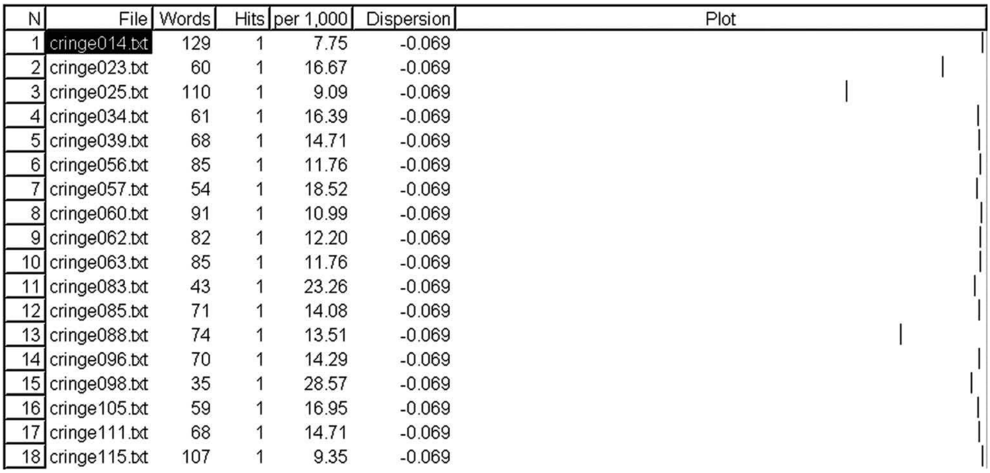
We have already seen that a dispersion plot suggests that the pattern *all of a sudden* frequently occurs midway in the texts that make up the corpus. Using the same tool, the distribution of *embarrassed* (both a keyword and the nucleus of two of the most frequent four-word clusters) is shown in Figure 20.2.

In other words, the vast majority of the occurrences of *embarrassed* occur at the very tail end of the text. By the same token, the majority (over 60 per cent) of occurrences of the phrase *one day* occur at the beginning of the text.

Working in this fashion – that is, plotting the distribution of keywords and high frequency clusters – the researcher starts to build up a composite picture of the generic structure of the text-type that makes up the corpus. Without detailing each step in the

Table 20.3 Four-word clusters with a frequency of $\geq$ 5 in the Cringe Text Corpus

N	*Four-word cluster*	*Freq.*
1	I was so embarrassed!’	12
2	in front of my	8
3	all of a sudden	7
4	was so embarrassed!’ I	6
5	in the middle of	6
6	front of my crush!	5
7	to go to the	5
8	friend and I were	5
9	right in front of	5
10	I had to go	5
11	was at school and	5
12	had to go to	5
13	the rest of the	5
14	‘I was at school	5
15	fell flat on my	5



N	File	Words	Hits	per 1,000	Dispersion	Plot
1	cringe014.txt	129	1	7.75	-0.069	
2	cringe023.txt	60	1	16.67	-0.069	
3	cringe025.txt	110	1	9.09	-0.069	
4	cringe034.txt	61	1	16.39	-0.069	
5	cringe039.txt	68	1	14.71	-0.069	
6	cringe056.txt	85	1	11.76	-0.069	
7	cringe057.txt	54	1	18.52	-0.069	
8	cringe060.txt	91	1	10.99	-0.069	
9	cringe062.txt	82	1	12.20	-0.069	
10	cringe063.txt	85	1	11.76	-0.069	
11	cringe083.txt	43	1	23.26	-0.069	
12	cringe085.txt	71	1	14.08	-0.069	
13	cringe088.txt	74	1	13.51	-0.069	
14	cringe096.txt	70	1	14.29	-0.069	
15	cringe098.txt	35	1	28.57	-0.069	
16	cringe105.txt	59	1	16.95	-0.069	
17	cringe111.txt	68	1	14.71	-0.069	
18	cringe115.txt	107	1	9.35	-0.069	

Figure 20.2 Dispersion plot for *embarassed* in the Cringe Texts Corpus

process, Table 20.4 displays some of the most frequent patterns, in the order in which they most commonly occurred, in the Cringe Text Corpus.

Cross-checking with the individual texts in the corpus, it is clear that they follow a narrative structure that shares characteristics with the structure described by Labov and Waletzky (1967), i.e.

> abstract – orientation – complication – evaluation – resolution – coda.

The abstract takes the form of a short title (not included in the corpus) such as *Dinner party disaster.* The orientation, which briefly mentions the circumstantial details against which the narrative action unfolds, frequently situates the event in the indefinite past (*one day* or *once*) and introduces a *hot* (or *cute*) *guy*, *older boy*, *senior*, etc., flagged as a key protagonist on first mention by the use of the determiner *this*. The complication is often signalled by a sentence beginning *all of a sudden/suddenly*. In seven of the texts further complications are introduced by the formulae *to make matters worse* or *even worse*. Finally, the evaluation includes derivations of *embarrass* or its synonyms.

Apart from the lack of a coda (which is in any case a minor, if not optional, element in narrative structures) the most significant difference between the teenage narratives and the Labov and Waletzky model is, in the former, the complete absence of a resolution. This finding is substantiated by the absence of any discourse markers, such as *fortunately*, *mercifully*, *happily*, etc., that might signal a resolution.

This raises the question: to what ends and in what contexts would narrators choose to purposefully tell stories at their own expense, leaving themselves in a state of unmitigated humiliation? Why introduce a complication and not resolve it? This in turn raises issues of age and gender: the stories were all written (allegedly) by teenage girls. Gender-based studies of language, such as Tannen 1994; Holmes 1995; and Coates 1996 and 2003, attest to the fact that the stories that women tell one another are often about personal misfortunes, their purpose being to elicit feelings of mutual empathy and to affirm their joint femininity. This is the 'point' of these narratives – and this is why they differ from the stories that men typically tell each other. As Coates (2003) notes: 'Self-disclosure is largely absent from men's narratives, but is a significant feature of the stories told by women to their friends' (2003: 118). And she adds, 'One of the rewards speakers get from self-disclosing is that fellow-speakers are likely to self-disclose in return. Reciprocal self-disclosure makes speakers feel supported by others, since the mirroring behaviour involved in reciprocal self-disclosure communicates understanding and empathy' (2003: 120). In short, story-telling is the way that women – and teenage girls – perform their gender.

A more critical analysis (e.g. Fairclough 1989; Lee 1992) might argue that such discursive practices maintain and reproduce asymmetrical power relations in society, and

Table 20.4 Frequency of some key phrases in the Cringe Text Corpus (n = 143)

N	*Pattern*	*Freq.*
1	one day/once	34
2	… this [really/totally] [hot/cute] [*male*]	10
3	all of a sudden/suddenly …	16
4	I was so embarrassed./ … it was so embarrassing.	25

that the teenage girls' magazines are complicit in a process of discursively positioning their readership as the helpless and disempowered objects of male derision. Again, a corpus search provides evidence of the 'objectification' of the protagonist: the pattern *laugh* at me*, for example, produces eleven instances, e.g. 'Everyone was staring and laughing at me.'

Writing of a related genre – first-person sex narratives in women's magazines – Caldas-Coulthard (1996) concludes,

> Sex narratives as cultural texts and discourses are responsible for maintaining a state of affairs which feminism has fought hard to change: inadequate and insecure women who, to have a voice, have to tell of their secret affairs even though they feel guilty.
>
> (1996: 269)

The same might be said about their daughters' 'cringe stories'.

At this point, we have moved from a discussion of text-level discourse analysis (*discourse^1*) into discourse-in-context (*discourse^2*), and, ultimately, discourse as social practice (*discourse^3*). Corpus analysis has identified surface-level features of the discourse(s) that inform the interpretative work at each stage. The procedure that has been outlined has attempted to show how a bottom-up approach, using a variety of corpus tools, can peel back successive layers of textual meaning. Of course, the same conclusions could just as well have been reached by a close reading of the actual texts, and without recourse to corpus tools at all. Nevertheless, the statistical data that a corpus approach delivers can serve to corroborate the findings of a more impressionistic approach, to confirm – or disconfirm – hunches, and to suggest new directions for further interrogation of the texts themselves. Schiffrin (1987), in arguing the case for the complementarity of quantitative and qualitative approaches in discourse studies, notes that 'quantitative analyses … depend on a great deal of qualitative description prior to counting (in order to empirically ground ones' categories) as well as *after* counting (statistical tendencies have to be interpreted as to what they reveal about causal relations)' (1987: 66). This cyclical alternation between counting and interpreting accurately characterises the application of corpus analysis to discourse.

5. What kind of data do you need to study discourse?

It goes without saying that discourse analysis requires texts – whole written texts, and, if not whole conversations, at least reasonably long stretches of (transcribed) talk. Since most discourse analysis focuses on textual features of specific text-types, a corpus that serves the needs of discourse analysis should consist of sufficient examples of these to provide generalisable data. But this does not mean it has to be enormous. For a start, 'in a collection of texts of similar type, the interactional processes and the contexts they take place in remain reasonably constant' (Partington 2004: 13). Consequently, 'specialised lexis and structures are likely to occur with more regular patterning and distribution, even with relatively small amounts of data' (O'Keeffe *et al.* 2007: 198).

The Cringe Text Corpus (see the previous section) is an example of the kind of corpus that targets a specific text-type and register. Another specific, small corpus that was relatively easily assembled, using texts available on the internet, is a 24,000-word corpus of abstracts from an academic journal. This consists of 139 texts, averaging 174 words each. It probably took about two hours to compile, cutting and pasting from the journal's website. (Obviously, publication of such a corpus, or of extracts from it, would require permission from the publishers.) A corpus this size is sufficient to provide information that can be reliably generalised for descriptive and pedagogical purposes.

For more rigorous research, some form of tagging – whether grammatical, semantic or phonological – is virtually obligatory. But, as Baker (2006) points out, tagging need not be exhaustive: 'Corpus builders need to think about what sort of research questions they intend to ask of their corpus, and then decide whether or not particular forms of tagging will be required' (2006: 42). If the focus, for example, is anaphoric reference, then only the referring expressions in the corpus, and their referents, need be tagged.

The advantage of a small, homogeneous corpus, such as one of journal abstracts, is that the context (of situation) can be precisely specified. As was noted in Section 3, if the study of discourse-as-language-in-context (*discourse^2)* is the aim, context information is essential. This is equally true for investigations into discourse-as-social-practice (*discourse^3*). As Mahlberg (2007) notes, in introducing her study of the local textual functions of the collocation *sustainable development:*

> The way in which an analysis of corpus data can be related to social situations depends on the information that is available on the origins and contexts of the texts. If the texts in a corpus are selected according to transparent criteria and information on their contexts is stored together with the texts, corpora can provide useful insights into meanings that are relevant to a society and indicative of the ways in which society creates itself.
>
> (Mahlberg 2007: 196)

In the future, the kinds of data that may be of increasing usefulness are those that support developments in the study of the emergent and ecological properties of language (Hopper 1998; Fill and Mühlhäusler 2001; Kramsch 2002; van Lier 2004), and especially the way that these properties are realised in discourse, both synchronically and diachronically. Developments in the application of complex systems theory to language acquisition and use (Ellis and Larsen-Freeman 2006; Larsen-Freeman and Cameron 2008) suggest we are experiencing the felicitous conjunction of two disciplines – corpus linguistics and psycholinguistics – that have, until now, tended to operate in parallel. After all, both are concerned with frequency effects (frequency of occurrence and frequency of exposure, respectively) and with the importance of usage (usage as data, and usage as performance). The happy alignment of the two fields is already influencing research: Ellis *et al.* (2008) report findings that show 'that formulaic expressions can be identified statistically from corpora of usage, and that native speakers and advanced ESL learners have become sensitive from their usage histories to these expressions so that they process them preferentially' (2008: 389), see also Lu (this volume). It is a short – but exciting – step from studying the processing of formulaic expressions preferentially to studying the processing of whole texts preferentially. Corpus evidence matched against data concerning the mental processing of texts may help reveal how patterns of text correlate with the way mental schemata evolve during comprehension and interaction. Likewise, research may show how the frequency of occurrence of particular discourses, and of variations in their texture, both influence and are influenced by the performance of these discourses – by individuals and across whole socio-cultural groups. As Larsen-Freeman and Cameron

(2008) point out, 'Conventional styles and registers of written text, such as "the small or classified ad" or "the academic essay", are emergent stabilities in the trajectory of social group written discourse. Genres are themselves dynamic and continue changing through use' (2008: 190). Corpus linguistics is well placed to track such changes.

Interdisciplinary collaboration coupled with technological advances in computation herald exciting developments in the field of *corpus discourse analysis*, and vindicate Hoey's (2005) claim

> that corpora are not just important for the study of the minutiae of language – they are central to a proper understanding of discourses as a whole, and that in turn means that there is no aspect of the teaching and learning of a language that can afford to ignore what corpus investigation can reveal.
>
> (Hoey 2005: 150)

Further reading

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum. (A practical introduction to applying corpus-based methodologies, such as collocations, concordances and dispersion plots, to the investigation of a range of different text types.)

Hoey, M., Mahlberg, M., Stubbs., M. and Teubert, W. (2007) *Text, Discourse and Corpora*. London: Continuum. (A collection of case studies, interleaved with insightful theoretical argument, that demonstrate the potential of corpus analysis to reveal aspects of textuality that might not otherwise be apparent.)

Partington, A., Morley, J. and Haarman, L. (eds) (2004) *Corpora and Discourse*. Bern: Peter Lang. (A collection of research papers that target a range of discourse areas, using corpus-based procedures, including discourse organisation, signposting and critical discourse.)

Scott, M. and Tribble, C. (2006) *Textual Patterns*. Amsterdam: John Benjamins. (A well-exemplified handbook on how to use corpus analysis tools (such as wordlists and keywords) in investigating the discourse features of a range of different registers.)

References

Aijmer, K. (1996) *Conversational Routines in English. Convention and Creativity*. London and New York: Longman.

Altenberg, B. (1990) 'Spoken English and the Dictionary', in J. Svartvik (ed.) *The London–Lund Corpus of Spoken English: Description and Research* (*Lund Studies in English 82*). Lund: Lund University Press, pp. 193–211.

Baker, P. (2006) *Using Corpora in Discourse Analysis*. London: Continuum.

Biber, D., Conrad, S. and Reppen, R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*. Harlow: Longman.

Biber, D., Conrad, S. and Cortes, V. (2004) '"*If you look at* …": Lexical Bundles in University Teaching and Textbooks', *Applied Linguistics* 25(3): 371–405.

Biber, D., Csomay, E., Jones, J. K. and Keck, C. (2004) 'Vocabulary-based Discourse Units in University Registers', in A. Partington, J. Morley and L. Haarman (eds) *Corpora and Discourse*. Bern: Peter Lang, pp. 23–40.

Bondi, M. (1999) *English across Genres: Language Variation in the Discourse of Economics*. Modena: Edizioni Il Fiorino.

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

Caldas-Coulthard, C. R. (1996) '"Women Who Pay for Sex. And Enjoy It": Transgression versus Morality in Women's Magazines', in C. R. Caldas-Coulthard and M. Coulthard (eds) *Texts and Practices: Readings in Critical Discourse Analysis*. London/New York: Routledge, pp. 250–70.

Carter, R. and McCarthy, M. (2006) *Cambridge Grammar of English*. Cambridge: Cambridge University Press.

Celce-Murcia, M. and Olshtain, E. (2000) *Discourse and Context in Language Teaching: A Guide for Language Teachers*. Cambridge: Cambridge University Press.

Coates, J. (1996) *Women Talk: Conversation between Women Friends*. Oxford: Blackwell.

——(2003) *Men Talk: Stories in the Making of Masculinities*. Oxford: Blackwell.

Coulthard, M. (1985) *An Introduction to Discourse Analysis*, new edition. London and New York: Longman.

Crookes, G. (1986) 'Towards A Validated Analysis of Scientific Text Structure', *Applied Linguistics* 7(1): 57–70.

Ellis, N. and Larsen-Freeman, D. (2006) 'Language Emergence: Implications for Applied Linguistics – Introduction to the Special Issue', *Applied Linguistics* 27(4): 558–89.

Ellis, N., Simpson-Vlach, R. and Maynard, C. (2008) 'Formulaic Language in Native and Second-Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL', *TESOL Quarterly* 42(3): 375–96.

Fairclough, N. (1989) *Language and Power*. London: Longman.

Fill, A. and Mühlhäusler, P. (eds) (2001) *The Ecolinguistics Reader: Language, Ecology and Environment*. London/New York: Continuum.

Fletcher, W. (2003/8) 'Phrases in English (PIE)', at www.usna.edu/LangStudy/PIE/ (accessed 7 January 2008).

Flowerdew, L. (2002) 'Corpus-based Analyses in EAP', in J. Flowerdew (ed.) *Academic Discourse*. London: Longman, pp. 95–114.

——(2003) 'A Combined Corpus and Systemic-functional Analysis of the Problem-Solution Pattern in a Student and Professional Corpus of Technical Writing', *TESOL Quarterly* 37(3): 489–512.

Gardner, R. (1998) 'Between Speaking and Listening: The Vocalisation of Understandings', *Applied Linguistics* 19: 204–24.

Ghadessy, M. (1988) 'The Language of Written Sports Commentary: Soccer – A Description', in M. Ghadessy (ed.) *Registers of Written English: Situational Factors and Linguistic Features*. London and New York: Pinter, pp. 17–51.

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

Hasan, R. (1989) 'The Texture of a Text', in M. A. K. Halliday and R. Hasan, *Language, Context, and Text: Aspects of Language in a Social-semiotic Perspective*, second edition. Oxford: Oxford University Press, pp. 70–96.

Hoey, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford University Press.

——(2005) *Lexical Priming: A New Theory of Words and Language*. London and New York: Routledge.

Hoey, M., Mahlberg, M., Stubbs., M. and Teubert, W. (2007) *Text, Discourse and Corpora*. London and New York: Continuum.

Holmes, J. (1995) *Women, Men and Politeness*. London: Longman.

Hopper, P. (1998) 'Emergent Grammar', in M. Tomasello (ed.) *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*. Mahwah, NJ: Lawrence Erlbaum, pp. 155–76.

Koteyko, N., Nerlich, B., Crawford, P. and Wright, N. (2008) '"Not Rocket Science" or "No Silver Bullet"? Media and Government Discourses about MRSA and Cleanliness', *Applied Linguistics* 29(2): 223–43.

Kramsch, C. (ed.) (2002) *Language Acquisition and Language Socialization: Ecological Perspectives*. London and New York: Continuum.

Kuiper, K. and Flindall, M. (2000) 'Social Rituals, Formulaic Speech and Small Talk at the Supermarket Checkout', in J. Coupland (ed.) *Small Talk*. Harlow: Pearson Education, pp. 183–208.

Labov, W. and Waletzky, J. (1967) 'Narrative Analysis: Oral Versions of Personal Experience', in J. Helm (ed.) *Essays in the Verbal and Visual Arts*. Seattle, WA: University of Washington Press, pp. 12–44.

Larsen-Freeman, D. and Cameron, L. (2008) *Complex Systems and Applied Linguistics*. Oxford: Oxford University Press.

Lee, D. (1992) *Competing Discourses: Perspective and Ideology in Language*. London/New York: Longman.

Mahlberg, M. (2007) 'Lexical Items in Discourse: Identifying Local Textual Functions of *sustainable development*', in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert, *Text, Discourse and Corpora,* London: Continuum, pp. 191–218.

Martin, J. R. and Rose, D. (2003) *Working with Discourse. Meaning beyond the Clause*. London: Continuum.

McCarthy, M. J. (1991) *Discourse Analysis for Language Teachers*. Cambridge: Cambridge University Press.

——(1998) *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

——(2000) 'Mutually Captive Audiences: Small Talk and the Genre of Close-Contact Service Encounters', in J. Coupland (ed.) *Small Talk*. Harlow: Pearson Education, pp. 84–109.

O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007) *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Partington, A. (2004) 'Corpora and Discourse, a Most Congruous Beast', in A. Partington, J. Morley and L. Haarman (eds) *Corpora and Discourse*. Bern: Peter Lang, pp. 11–20.

Pulcini, V. and Furiassi, C. (2004) 'Spoken Interaction and Discourse Markers in a Corpus of Learner English', in A. Partington, J. Morley and L. Haarman (eds) *Corpora and Discourse*. Bern: Peter Lang, pp. 107–24.

Schauer, G. and Adolphs, S. (2006) 'Expressions of Gratitude in Corpus and DCT Data: Vocabulary, Formulaic Sequences and Pedagogy', *System* 34: 119–34.

Schiffrin, D. (1987) *Discourse Markers*. Cambridge: Cambridge University Press.

——(1994) *Approaches to Discourse*. Oxford: Blackwell.

Scott, M. (1997) *WordSmith Tools Manual*. Oxford: Oxford University Press.

——(2008) *WordSmith Tools* version 5. Liverpool: Lexical Analysis Software.

Scott, M. and Tribble, C. (2006) *Textual Patterns*. Amsterdam: John Benjamins.

Sinclair, J. (ed.) (1990) *Collins COBUILD English Grammar*. London: Collins.

Stenström, A.-B. (1994) *An Introduction to Spoken Interaction*. London: Longman.

Stubbs, M. (1994) 'Grammar, Text, and Ideology: Computer-Assisted Methods in the Linguistics of Representation', *Applied Linguistics* 15(2): 201–23.

——(1996) *Text and Corpus Analysis*. Oxford: Blackwell.

——(2001a) *Words and Phrases: Corpus Studies in Lexical Semantics*. Oxford: Blackwell.

——(2001b) 'Text, Corpora, and Problems of Interpretation: A Response to Widdowson', *Applied Linguistics* 22(2): 149–72.

——(2005) 'Conrad in the Computer: Examples of Quantitative Stylistic Methods', *Language and Literature* 14(1): 5–24.

——(2007) 'Quantitative Data on Multi-word Sequences in English: The Case of the Word *world*', in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert (eds) *Text, Discourse and Corpora*. London: Continuum, pp. 163–90.

Swales, J. M. (1990) *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, J. M. and Najjar, H. (1987) 'The Writing of Research Article Introductions', *Written Communication* 4: 175–92.

Tannen, D. (1994) *Gender and Discourse*. New York: Oxford University Press.

Tao, H. (2003) 'Turn Initiators in Spoken English: A Corpus Based Approach to Interaction and Grammar', in P. Leistyna and C. Meier (eds) *Corpus Analysis: Language Structure and Language Use*. Amsterdam: Rodopi, pp. 187–207.

Teubert, W. (2000) 'A Province of a Federal Superstate, Ruled by an Unelected Bureaucracy: Keywords of the Eurosceptic Discourse in Britain', in A. Musolff, C. Good, P. Points and R. Wittlinger (eds) *Attitudes Towards Europe: Language in the Unification Process*. Aldershot: Ashgate, pp. 45–86.

——(2007) '*Natural* and *Human Rights, Work* and *Property* in the Discourse of Catholic Social Doctrine', in M. Hoey, M. Mahlberg, M. Stubbs and W. Teubert, *Text, Discourse and Corpora*. London: Continuum, pp. 89–126.

van Lier, L. (2004) *The Ecology and Semiotics of Language Learning: A Sociocultural Perspective*. Dordrecht: Kluwer.

Widdowson, H. G. (2004) *Text, Context, Pretext: Critical Issues in Discourse Analysis*. Oxford: Blackwell.

——(2007) *Discourse Analysis*. Oxford: Oxford University Press.

Yates, S. J. (1996) 'Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study', in S. Herring (ed.) *Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam: John Benjamins, pp. 30–46.