This is chapter 1 of *Text and Corpus Analysis* which was published by Blackwell in 1996.

TEXT AND CORPUS ANALYSIS:
COMPUTER-ASSISTED STUDIES OF LANGUAGE AND CULTURE

MICHAEL STUBBS

CHAPTER 1.

TEXTS AND TEXT TYPES

> Attested language text duly recorded is in the focus of attention for the linguist. (Firth 1957b: 29.)

The main question discussed in this book is as follows:

> How can an analysis of the patterns of words and grammar in a text contribute to an understanding of the meaning of the text?

I discuss traditions of text analysis in (mainly) British linguistics, and computer-assisted methods of text and corpus analysis. And I provide analyses of several shorter and longer texts, and also of patterns of language across millions of words of text corpora. The main emphasis is on the analysis of attested, naturally occurring textual data. An exclusive concentration on the text alone is, however, not an adequate basis for text interpretation, and this introductory chapter discusses the appropriate balance for text analysis, between the text itself and its production and reception.

Such analyses of attested texts are important for both practical and theoretical reasons, and the main reasons can be easily and briefly stated. First, there are many areas of the everyday world in which the systematic and critical interpretation of texts is an important social skill. One striking example occurs in courtrooms, where very important decisions, by juries as well as lawyers, can depend on such interpretations. Second, much of theoretical linguistics over the past fifty years has been based on the study of isolated sentences. This approach has some severe limitations and it is important to show what can be learned by studying patterns of language across texts and corpora.

1.1. THE DATA

By text, I mean an instance of language in use, either spoken or written: a piece of language behaviour which has occurred naturally, without the intervention of the

linguist. This excludes examples of language which have been invented by a linguist merely to illustrate a point in a linguistic theory. Examples of real instances of language in use, might include: a conversation, a lecture, a sermon, an advert, a recipe, a newspaper article, a scientific research paper, a novel, a school textbook, and so on. The list is open-ended, and probably endless. In chapter 2, I discuss in more detail the advantages of basing linguistic description and theory on attested instances of language use.

One brief point of terminology. There is considerable variation in how terms such as *text* and *discourse* are used in linguistics. Sometimes this terminological variation signals important conceptual distinctions, but often it does not, and terminological debates are usually of little interest. These distinctions in terminology and concept will only occasionally be relevant for my argument, and when they are, I will draw attention to them (e.g. in chapter 7.2).

It is an important feature of the book that all the data which are analysed in any detail are attested in this sense. Where this is not absolutely clear from the context, I will use the following conventions (see also section on Data Conventions and Terminology) to mark the source of examples:

[A]	actual, authentic, attested data
[M]	modified data
[I]	invented, intuitive, introspective data.

Almost all examples in the book are [A]. Occasionally, it is convenient to modify (for example, to abbreviate) or to invent an example, and I will always warn the reader when I have done this.

Possibly the most central problem in contemporary linguistics is: What counts as data? This question has major implications for descriptive and theoretical linguistics. What precisely is the evidence on which theories of syntax and semantics are based? But questions of what counts as evidence, corroboration or proof are also important in applied linguistics, and I will discuss in detail texts from the mass media, education and courtrooms. In such cases, linguists have a responsibility to say how confident they are that their analyses are correct 'beyond reasonable doubt'.

1.2. THE ORGANIZATION OF THE BOOK

Chapters 1 to 3 discuss the concepts of text, text type and genre, with particular reference to the British neo-Firthian tradition of linguistics, as it has been developed by Halliday and Sinclair. This tradition has led to major grammars and dictionaries of English, and to significant advances in methods of computer-assisted text and corpus analysis.

Chapters 4 to 8 provide analyses of texts and text corpora. One descriptive focus of these chapters is texts and text types which are public and/or authoritative: two widely circulated speeches (chapter 4), the summing-up by a judge in a criminal trial (chapter 5), two school textbooks (chapter 6), statements by public figures (chapter 7), and various corpus data (chapters 7 and 8). These chapters progress from little things to big

things: from short texts (of a few hundred words) to long texts (whole books of tens of thousands of words) from a comparison of one text with another, to a comparison of patterns in texts with patterns in corpora and from single texts to text corpora: that is, collections of texts of millions of words in length.

This progression is important for two reasons. First, it should facilitate the use of the book in teaching. These chapters introduce various methods of computer-assisted text and corpus analysis: in particular, the use of concordances for studying patterns of language use. Later chapters also introduce other methods for studying the most frequent and characteristic syntactic constructions and lexical collocations in which words occur.

Second, it facilitates a discussion of criteria for text analysis. These include: the need to analyse not only short text fragments, but also whole long texts; and the need for the stylistic analysis of individual texts to be based on comparisons with other texts and with corpus data which represent (however imperfectly) the language. The main descriptive and theoretical arguments in the book concern findings about language which could not conceivably result from intuitive data in the form of invented sentences, but which require large quantities of corpus data. (See especially chapters 2, 7, 8.)

1.3. A SIMPLE MODEL: TEXTS, PRODUCTION AND RECEPTION

This opening chapter makes some introductory points about the relations between texts, text types and social institutions, and gives some examples of different studies. A problem which crops up in various forms is: Where is the meaning of a text located? Is the meaning inside the text itself? Or inside the mind of the person who makes sense of it? Or is it in the speech community somewhere? – perhaps in the form of a consensus interpretation on which we could all agree?

Suppose I receive a legal document, which I read, but do not fully understand. I might ring up a friend who is a lawyer and read a problematic sentence over the phone. My friend might then explain to me what this sentence means. I will have read it, but the lawyer will have interpreted it. So was the meaning in the text? Possibly, although it seems to require the knowledge of the lawyer to get the meaning out. And note the ambiguity of the word *read*. It sometimes refers merely to the words on the page: I have read a sentence aloud over the telephone. And it sometimes means "read and understand", as when I say that I have read a good book on holiday.

Here is a very simple model of the relation between text and context. In common sense terms, it seems clear that the meaning of a text depends on at least three things: the language of the text itself, who produced it, and who is responding to it. By the language itself I mean the words actually spoken or written, and their patterns of lexical collocations and of syntactic and rhetorical structures. Some meanings are created by the words themselves and their observable interrelationships, and several chapters below are concerned with analysing such patterns. But some meanings depend on our knowledge of the point of view of the author. We may interpret things quite differently, depending on when and where and by whom the language was produced, for example, depending on the status or authority of the speaker or writer: a person in the street, an

expert, a government minister, a government committee, an author of a school textbook, a judge summing up in court, and so on. And some meanings are brought to the text by readers or listeners: according to their specialist knowledge, their cultural assumptions, or their familiarity with other related texts. Readers and listeners also have different points of view, and respond to texts in different ways.

The text-production-reception model emphasizes immediately that meanings and interpretations are mediated by social institutions. It is difficult to discuss examples of language, say from a textbook or a judicial summing-up, without referring to the status of speakers and listeners in institutions such as education and the law.

This simple view of different factors in the meaning of texts is reflected in the history of literary criticism over the last hundred years or so. (Eagleton, 1983, provides a good discussion of this history.) There was a period when the main focus was on the author: his or her biography and social background were felt to be crucial to an understanding of a text. This tradition is still represented in the essays of literary reference books. An essay on Wordsworth, for example, might combine comments on the themes of his poetry, with a brief account of his life and references to contemporary historical events, such as the French Revolution.

In the 1920s, various approaches shifted the focus to the text itself, the words on the page. Critics emphasized the formal aspects of texts, and saw literature as a particular organization of language, sometimes even downgrading content in favour of form. Close critical reading of the text, sometimes in deliberately maintained ignorance of the author (since this might influence interpretations) was recommended by I. A. Richards (1929), in his approach to practical criticism. As part of the view which treated literary works as objects, independent of their historical and social contexts of production, the intentional fallacy refers to the argument that we do not (or cannot or need not) know the intention of an author in order to interpret his or her text. The text is autonomous; the author is irrelevant to its interpretation.

But the practical criticism position already implies that readers' interpretations can be influenced by what they know: that is, by things outside the text. And more recently, the focus has been on the reader. Reader response theory emphasizes that meanings are constructed by active readers: they are not derivable from texts in a passive way. This position is represented in influential work by Iser (1974) and Fish (1980).

And views are still shifting on the relation between text, author and reader. After more extreme views on the 'death of the author', literary critics have begun more recently to argue again that an author's biography may, after all, shed light on the text, and many well publicized literary biographies have recently appeared. In addition, the concept of authorship, and dependent concepts such as copyright and plagiarism, are themselves historically variable. In fact, the concept of an individual creative author is a relatively recent historical invention: largely a product of Romanticism. We do not know the author of some famous Middle English literary texts (such as *Sir Gawain and the Green Knight*, written some time before 1400): this seems not to have been important enough to record. And most texts in the modern world, such as government reports, advertisements, and even many newspaper articles, have no identifiable author. They are anonymous, or have multiple authors or ghost authors.

These shifts in the fashion and theory of literary criticism, along with historical shifts in how authors are regarded, show the constantly changing relation between authorship and authority. And they all seem to indicate that the meaning of a text is not in any single place. An exclusive reliance on any one of these sources of meaning is not adequate, since all three contribute to our interpretations of texts. This simple model, which relates the text itself, speakers or writers, and listeners or readers, draws attention to relations which are essential in text analysis. A balanced approach must take into account both product and process: not only the text itself, but also its production and reception.

Usually, given a text, it is possible to infer quite a lot about who produced it for whom; and given information about author and audience, it is often possible to predict quite a lot about what kind of thing the author is likely to say in what way. We can do this because even creative literary authors are not entirely free to express things as they wish. An author's freedom is always constrained by how other speakers use the language and by accepted ways of speaking and writing. The concepts of prediction and probability are central to the view of language which I develop below.

Most chapters in this book discuss texts with reference both to their linguistic patterning and also to their contexts of production and reception. But it is not possible to study everything at once, and some chapters focus largely on the text itself, whereas others emphasize relations between a text and its social and political background.

I will also discuss patterns which are in the language. Literary and linguistic theory have also emphasized that it is not only texts which can be anonymous. The language itself is anonymous. It is never the property of individual speakers. Individuals have to take over ways of speaking from previous generations of the speech community (see chapter 3.4). Along with individual words and grammatical rules, speakers acquire routine expressions, common ways of formulating things and collocations, which encode commonly accepted ideas (see chapters 4 and 7).

1.4. CONTENT, AUTHOR AND AUDIENCE

This simple three-part model for thinking about language and interpretation provides a route into several aspects of text analysis. The model draws attention both to the expression of content, and also to the expression of relations between author and audience. For example, in analysing a school textbook, it is natural for the focus to be on the content of the text. This is likely to be what the author was concentrating on: a school textbook is about some subject, such as mathematics or geography. But it may be important to pay attention also to interpersonal meaning: does the author admit to being a fallible human being, or claim unassailable authority for the truth of what is written? (A detailed example is given in chapter 6.10.)

The model also emphasizes that texts do not have absolute and unchanging interpretations, but are interpreted differently in different historical periods, and in different cultural contexts. Indeed a major source of misinterpretation is when texts are read outside a specialist context. (Chapter 7.3 gives examples of what can happen to

linguistic arguments when they are recontextualized in educational and political debates.)

Each genre of text, such as Bible translations, novels, reading primers, textbooks, and newspapers, has served social and cultural functions. The authority of textbooks is related to the view that the meaning is in the text. But very different views of knowledge are implied by different relations between text and audience: texts can be dictated, learned by rote, read silently in private or aloud in public, discussed in groups, and so on. The most basic distinction is perhaps between an unmediated reading and an expert interpretation.

It is worth remembering that the Reformation turned on debates about who could interpret what. When Luther posted his theses on the door of a church in Wittenberg in 1517, he was challenging the Papal authority to interpret the scriptures, arguing that everyone should read the Bible for themselves. The translation of the Bible from Latin into German was a challenge to the mediation of the text by the organized Church. Luther's view was that people should study the text itself, not rely on the claimed authority of other people's interpretations: 'The meaning of the Scripture depends, not upon the doctrine of the Church, but on a deeper reading of the text.' He was thereby aiming to change the relations between text, social institution and authoritative interpretation. Such questions of text, audience and interpretation are not, as they say, merely of academic interest (not merely of interest to literary theory, for example). Changed views about their relations have been at the root of religious and social revolutions. (And different religions have taken very different views about text and translation.)

A model which emphasizes the relations amongst text, production and reception points the way to a critical model of text interpretation, in which listeners and readers do not simply accept received opinion, but can identify and discuss cases where evidence is given for views versus cases of personal opinion, unsupported assertion, innuendo or bias, and can better understand the differences between what the words mean, what the writer/speaker means, and what the reader/listener understands.

In Shakespeare's play *Julius Caesar* (act 3, ii), after Caesar's murder by Brutus and others, Antonius gives his famous speech which starts *Friends, Romans, Countrymen, lend me your ears*. He repeats several times Brutus' claim that Caesar was ambitious, and says, also several times, that *Brutus is an honourable man*. The words mean one thing, the speaker means something else, and the listener's understanding of the words changes as Antonius repeats them throughout his speech. The words have a micro-history in the text (see chapter 4.5). Nowadays this Shakespearian line is so well known that it can itself be quoted to draw attention to such ambiguous and shifting meanings. I recently overheard a conversation in which a speaker was criticising a colleague, but concluded, *But he is an honourable man*. To which the response was: *Brutus was an honourable man.* Words and fixed phrases also have a history in the language (see chapter 7).

A method of teaching students, including school pupils, to interpret the points of view from which texts are written has many educational implications. This will not be a

central focus in this book, although I will draw attention to educational implications of the argument when this is particularly striking. (See chapters 4.9 and 6.)

1.5. TEXTS AND TEXT TYPES

In chapters 4 to 6, I will analyse examples of different texts which represent different configurations of text type, author and audience:

- two speeches: produced by a public figure, and interpreted by a wide range of readers including children
- summing-up in a criminal trial: produced by the judge, and interpreted by lawyers, the defendant and members of the jury
- two school textbooks: produced by authors, and interpreted by pupils and teachers.

I will also give briefer examples (chapter 7.3) of what can happen when ideas about language and education are discussed in the mass media by public figures. Other corpus data which I analyse comprise samples from many other text types. These various analyses will show how messages are conveyed: not only explicitly, by words themselves, but also implicitly, by lexical and syntactic patterning.

Again terms are variable. Some authors distinguish between text type and genre: I will not. Text types or genres are events which define the culture. They are conventional ways of expressing meanings: purposeful, goal-directed language activities, socially recognised text types, which form patterns of meaning in the social world (Kress 1989). An essential concept is purpose for audience. Text types also imply different ways of producing, distributing and consuming texts (Fairclough 1992: 71). They may be produced individually (a personal letter) or collectively (a committee report), the author may be named (a novel) or anonymous (a television ad); they may be published, xeroxed, faxed; they may be read once and thrown away, or read repeatedly (a work of literature), or learned by heart by successive generations (famous poems, or many religious texts); they may be private or public.

Genre, as a traditional category in literary studies, includes short story, novel, play, autobiography, diary, sonnet, epic and fable. It therefore fits naturally into a study of the aesthetic functions of language. However the concept is also relevant to broader forms of cultural analysis. Popular culture provides examples such as science fiction, detective fiction, romance, comic strip and western. Some genres have developed for use in television and radio: documentary, soap opera, phone-in, panel discussion, news broadcast, travel programme, light entertainment and quiz show. These examples emphasize that genres change and evolve. Relatively recent genres include the blues, the presidential press conference, the television game show, the music video (Swales 1990: 33-35). And the concept applies equally to everyday uses of spoken and written language, such as joke, story, chat, gossip, song, sermon, argument, debate, committee meeting, instructions and signs. When using various corpora of contemporary English (such as LOB and the Longman-Lancaster corpus), it is important to bear in mind that many such genres are not represented: diaries, personal letters, business correspondence (a huge genre), company reports, writing for teenagers, and so on.

A genre which has become prominent in modern life consists of guides and manuals: books and articles offering practical, easy-to-read, concise, comprehensive advice on various subjects. Any bookshop contains many such books and magazines which provide advice on self-help, self-management, self-realization, self-knowledge, life-style and efficiency. Common topics include: the body (health, fitness, diet, etc.); personal relations (in marriage and divorce, in child-care, at work, etc.); the home (buying and selling houses, moving house, do-it-yourself, gardening, cooking, etc.). One can buy a book of advice on almost anything: how to cure migraine, how to set up a business, how to avoid paying too much tax, how to cross the Australian outback in a four-wheel drive vehicle. Whole magazines are devoted to topics such as health and diet, personal relations, work and travel. Giddens (1991) has pointed to this vast literature as a feature of modernity. There is no clear dividing line between the technical or expert and the popular, between social science and guides to life, and some writers produce both (for example, Deborah Tannen writes both on discourse analysis and on communication problems in marriage). Giddens sees this ambiguity as an important part of the reflexivity of modernity.

Genres can also go out of fashion or become less useful. The essay form no longer has the prestige it had in the eighteenth century when Addison and Steele were writing for the *Tatler* and the *Spectator*, though a few periodicals still publish essays on general political and social issues. The pastoral and the epic poem are today dead as literary genres. And the role of genres in society changes: think of political satire, proverbs, traditional tales, sermons and funeral orations.

The concept of text type is clear enough in general, but although many categorizations have been proposed, none is comprehensive or generally accepted. There is no implication that such genres are categories with neatly defined boundaries; although the focal members of genres are usually easy to identify. Genres can, for example, be combined. The important point is not knowing in some mechanical way which genre an example fits into: but knowing how the category can make a difference to the way in which it is interpreted. The emphasis in all such study is on the conventions which govern the uses and meanings of language. The ability to identify and compare different genres contributes to our ability to understand them. Misunderstandings, and even dislike, of texts of different kinds can often be due to a lack of understanding of the different conventions involved.

1.6. TEXT TYPES AND INSTITUTIONS

Social institutions and text types are mutually defining. An example is provided by the public media – radio, television, newspapers – where new text types have developed with the media, and now help to define them. Some valuable historical studies are available.

Püschel (1991a, b) discusses the development of new types of newspaper article which arose in Germany in the nineteenth century. He discusses the shift from merely reporting that events had taken place, to more detailed commentary on events, and to the formation of public opinion through rational public political debate. This involved the differentiation of genres such as short news items and longer reports, on-the-spot and eye-witness reports, background coverage, commentary, and leading editorial articles. It

also involved page layout, which helped readers to distinguish between different types of item such as readers' letters. The letter to the editor was itself a genre which had to be invented and developed.

Scannell (1986) discusses the development of documentary programmes on BBC radio from the 1920s to the 1940s. The BBC had an explicit policy of finding new forms of communication to transmit socially contentious issues to a new mass audience. They broke with established literary modes, and attempted to record 'what actually happened', and to allow people and events to 'speak for themselves'. They saw the audience as being within family groups in the home, for which intimate, informal styles were appropriate. They deliberately contrasted the points of view of officialdom and of working class people. This all represented attempts to reconstruct the relations between the institution of the BBC, the subject matter of programmes, the people who had access to the microphone, and the audience. There were also sharp differences between national broadcasting from the south of England (in London) and regional programmes produced in the north (in Manchester). A new concept of a general public was developed, with a more local working class public for regional programmes. And new genres were developed: documentaries, features, social reportage, eye-witness reports, dramatized reconstructions, studio discussions, dialect plays. The sound of a live audience for the first time in a radio programme was heard as a major innovation. Interviewing was a new technique and had to be invented. News programmes only existed from 1934. News interviews were not used until after the War.

Scannell (1986) documents the changing relations between institution, genre, subject matter, speakers and audience, and the ways in which a new institution and new genres were developed in tandem. (He does not provide any textual or corpus analysis to ground his analysis, though to be fair, much broadcasting of the period was live, and there are no surviving recordings to analyse!)

Another set of links between genres and institutions can be seen if we look at professions which specialize in discourse (Bernstein 1990: 135ff), such as priests, scientists, lawyers, teachers and administrators. Priests, for example, are specialists in particular types of discourse such as prayers, sermons, confessionals, baptisms, excommunications and exorcisms! And they are experts in the interpretation of particular types of written texts: scriptures and related genres. Some professions write and sell texts: advertisers and journalists. Some professions reproduce and mediate texts: teachers, priests, actors. A public relations agency is paid to transform one text into another, to create new texts from different sources, in order to try to change people's beliefs, persuade them of a point of view, or alter the actions of politicians.

Such transformations can lead to long intertextual sequences. For example, scientists write research reports and articles on the destruction of the ozone layer. These technical texts are interpreted by experts in other fields (for example, those drafting industrial legislation, or environmental organizations), and by journalists. They are also rewritten by public relations agencies who prepare position papers and speeches for industrial clients opposed to the legislation. A detailed case study of the mediation of knowledge in this area by different interest groups is provided by Gerbig (1993).

In addition, there are several relatively new professions which not only specialize in discourse, but whose discourse is designed to regulate psychological and social relations. These are the professions to do with counselling and guidance: child guidance and marriage guidance, personnel officer, social worker, psychoanalyst, psychotherapist, public relations agency. The growth of such professions since the eighteenth century, and the kinds of power they have, based on the new kinds of knowledge developed within the social and behavioural sciences, is one of Foucault's (1980) main themes. The colonization of new areas of life by the discourse of counselling is discussed by Fairclough (1992). (See chapter 7.10 for the importance of cultural key words such as *expert* and *professional*.)

1.7. THREE STUDIES

When we look at published studies of texts and text types, we find all conceivable combinations of focus on text, production and reception. Studies focus variously on: the text itself, patterns of words or grammar; the relations between one text and another, patterns that we recognize from other texts which we have heard or read; text corpora, patterns across large collections of texts; and the context, relations between text, author and audience, or the broader social and political context of the text.

1.7.1. Advertising

There have been many studies of the language of advertising: two worth consulting are Leech (1966) and Myers (1994).

Cook (1992) provides a highly readable, and often witty, study of advertisements in British magazines and television. He discusses the genre in general, different types of ads (for products, charities, health campaigns, and so on), and changes in advertising fashion from the 1950s to the 1990s. His analytic method is line-by-line commentary on the short texts and their visual and musical components, and his observations on individual ads are invariably perceptive. However, ultimately, the method is simply that of confident personal literary judgement. Thus a poem in one ad is condemned (p. 123) as 'outstandingly banal and clumsy ... a bad ad as well as a bad poem'.

Cook mentions several methodological issues, but does not draw the consequences in his own analyses. He emphasizes (p. 199) that people have 'vast and daily experience' of advertisements and that 'each new ad is encountered through knowledge of thousands of earlier ads'. But the book is not based on a defined corpus of data, and he can only appeal to our intuitions that a particular television ad is 'fairly ordinary' (p. 38) or that other examples are 'prototypical' (p. 8). He admits (p. 11) that he ignores the numerically largest sub-type in the genre, small ads in newspapers.

He proposes sampling a wide range of data: it is (p. 11) 'most instructive to look at ads which occupy extreme points' in the genre. But he looks in detail at only one such point, the allusion and verbal play in soft-sell ads. Here he picks his own favourites (which is what all literary critics do), concentrating on memorable or famous examples, but does not analyse the majority(?) which either merely provide useful information and/or are just banal. His examples are mainly of the type where buying a brand of chewing gum, four-wheel drive car, jeans or instant coffee makes you intelligent, good-looking, loved

by your family or desired by the opposite sex (in that order?). He mentions many other types (for example, for political parties, famine relief, anti-drink/drive campaigns), but does not analyse examples. And he does not discuss other extreme points, such as the marketing of the American President, or radio travel programmes which are extended advertising for package holidays.

Cook's study contains precise and interesting commentary on individual short texts, and his analyses draw attention to features I would certainly not have noticed. However, given its basis in personal commentary, and the failure to specify the corpus, it is difficult to see how the study could be replicated. He emphasizes that people are not cultural dopes who believe everything in soap operas or ads. But his only audience research refers (pp. 106, 112) to 'an informal survey of forty young adults' (some of his students?).

Cook mentions all the following themes, but has no explicit theory to explain:

- how an individual text can have meaning only as a sample of an enormously large body of texts
- how typical or representative texts can be identified
- how the relation between text and audience could be studied.

1.7.2. Newspapers

There are also many analyses of newspaper language: Bell (1984, 1991) and Fowler (1991b) are particularly useful studies.

In a study which is very different from Cook's in style and method, Jucker (1992) provides a detailed analysis of a sample of British newspapers, based on a well defined sub-set of English grammar, the structure and complexity of noun phrases (NPs), and on a well defined corpus. The corpus comprises samples, from a few days in 1987-88, of all eleven British national daily newspapers, up-, middle- and down-market. Based on circulation figures, he provides empirical evidence of their audiences. He then makes detailed comparisons across the newspapers, and also across different sections (genres?) within them (UK and foreign news, business, arts and sports). The 371 articles studied provide 43,000 NPs. He describes the syntax of the NPs, with statistics on their distribution in different newspaper types and sections, to give a careful grammar of NPs, with probabilities attached, on the basis of a precisely defined, though narrow, sample of real written data.

However, what is missing is a semantic analysis. This is not an oversight, of course, but a deliberate part of the methodology. Yet without a discussion of meaning, it is difficult to know what to make of the detailed descriptive statistics. As Jucker himself admits, it is 'a complex step to reach any conclusions about the stylistic aims of specific newspapers and newspaper sections' (p. 148). Three (out of many) clear findings are as follows. The average number of modifiers per 1,000 NPs distinguishes the newspaper types: up-markets have more (p. 108). NP-appositions stratify the three newspaper types more distinctly than any other structural property of NPs (pp. 107). And the down-market papers always omit the determiner (a, the) in NP-appositions (p. 230), e.g. *Royal*

relative Katie Baring, left-wing firebrand Derek Hatton. But Jucker provides little explanation of what these differences might mean.

An interpretation might proceed as follows. A language always provides different ways of describing a common world, and Jucker shows that different newspapers have different ways of referring to the same people, things and events. NPs provide ways of referring to people (*Tory health fanatic Edwina Currie, a splendidly gloomy Vladimir*), things (*Perth jail, a 130 mph silver X-registered BMW coupé*), and events (*a devastating onslaught, the tragic news*). (These are attested examples from the book.) But what is missing is a discussion of what these nomenclatures imply about how the world is encoded. For example, there are many illustrations of the elaborate taxonomies used for professions, titles and personal descriptions (*father, husband, billionaire, minister, gun fanatic, student nurse, crackpot, curvacious cutie*, etc., etc.), but no analysis of how such taxonomies condense and classify experience. (See chapter 7 on cultural keywords, and chapter 8.8.1 on words such as cutie.)

Two types of information, which would permit such an interpretation, are missing. First, Jucker provides no semantic information on the NPs. Most examples refer to people (presumably in line with the well known tendency of the press to personalize stories), but he gives no statistics on what the 43,000 NPs most often refer to. Events? things? people? women? men? Presumably women and men are categorized with different modifiers: name? profession? appearance? With such semantic information we could better interpret the kind of world which is being constructed. Second, the corpus is not in fact a sample of texts, but the list of NPs, extracted and studied independently of their co-text. Jucker studies the NPs in the context of type of publication (up- and down-market) and text type (e.g. sports versus arts section), but he ignores the data needed to interpret the isolated grammatical constructions. Since he ignores the intervening layers of sentence and text, he cannot explain how such NPs are used to construct arguments.

Jucker's careful quantitative study of the frequency of syntactic constructions across text types provides normative data which will be very helpful in future comparative studies of the meaning of syntactic patterns. But Jucker's study stops short of several questions:

- there is analysis of the distribution of grammatical structures across texts, but no analysis of their meanings
- there is no interpretation of syntactic features in their co-text
- there are statistics of features of the corpus, without any study of the constituent, individual texts.

1.7.3. Scientific research articles

There are also many analyses of scientific English. The study by Halliday and Martin (1993) is particularly relevant to chapter 6 below.

Swales (1990) provides a study of scientific research articles, which is an important contribution to understanding academic discourse and also to genre theory. It discusses major approaches to genre in folklore studies, literary studies, linguistics and rhetoric. A genre is seen as a class of communicative events, with shared communicative purposes. Discourse communities share agreed public goals and mechanisms of communication,

and possess one or more genres. Competence in the genre(s) is required for membership of the community.

Swales sees science as comprising various international discourse communities, whose central communicative mechanism is the research article, plus associated genres such as abstract, research presentation, grant proposal and thesis. He points out (p. 95) that the research article is a huge genre: there are perhaps 100 thousand research journals in the world, in science, technology and other subjects, publishing perhaps 5 million articles per year. Some journals are slim and intermittent, but others are huge: *Physical Review* published 30 million words in 1980.

He provides an account of the historical development of scientific English out of personal letters amongst scientists from the mid-1600s onwards. The genre was consciously developed by scientists who required ways of expressing generally accepted knowledge about experimental matters of fact. He then summarizes a great deal of work on the syntax and discourse structures preferred in research articles, including: tenses, modals, nominals, actives and passives, commitment and hedging, topic sentences, text structures (e.g. problem-solution, introduction-methods-results- discussion), authorial comments, citation patterns, and information density. He refuses any simplistic clichés, such as scientific language uses an impersonal nominalized style.

Swales emphasizes the role of language in science, the 'sheer importance of the writing' (p. 127), and talks of the documentary world (p. 122) of science, in which articles are products and in which the discussion of written formulations is a major activity. Swales sees reality as linguistically constructed. He discusses the 'construction of research articles', the 'long process of rhetorical construction' (p. 121) from experimental work via laboratory reports (a very different genre) to publication, and the 'creation of facts' (p. 112) through this 'complexly distanced' (p. 175) rhetorical construction, where the aim of the rhetoric is to convince the reader that there is no rhetoric. He seems unhappy about the subjectivist implications of this constructionist view (p. 122), although he has a concrete and institutional view of process and product. Knowledge is produced by scientists working in institutions. The main product of this knowledge-manufacturing industry is research articles (pp. 95, 125), which are distributed along communicative networks: they may be selected for inclusion in journals, further selected by abstracting services, requested as reprints by other scientists, mediated by science journalists in popularized accounts, and required by universities for promotion. Swales' discussion of these gate-keeping policies provides much empirical evidence for a theory of the relations between texts, knowledge and communicative networks.

One unclear point is his use of the term 'translation' to refer to journalists mediating original research articles and turning them into popularizations. It is possible to translate between languages, dialects or styles, where it is meaningful, if approximate, to talk of the same content being conveyed in different forms. But when the shift is between genres (technical and popular), then it is not the same content in different words, but different content. Scientists produce the original texts. These are then both modified and simplified, and also recontextualized by being put into contact with different texts: no longer other research texts, but, say, news items about politics. (On the recontextualization of knowledge, see chapter 6 on school textbooks, and chapter 7.3 on public figures making statements about education.)

An omission in the argument is as follows. Swales points to the great variation in the scale and level of linguistic analysis in work on research articles. Some studies are based on a single, sometimes atypical, article; other studies give no indication of what data they are based on! But Swales himself makes no attempt to compare this special genre with academic English in general, or with written and spoken English more widely. Any study of genres must be located in a description of variation in the language overall. Swales reviews many linguistic features which have been found in scientific research articles, but does not relate this to a theory of variation in English. As Biber (1988: 178) has shown, there is a wide range of variation within academic prose, and therefore no individual text represents this genre adequately. Furthermore, the range of texts within natural science is even more striking, depending on the sub-discipline, say engineering versus biology (Biber, 1988: 193).

Swales could also have developed his argument with reference to Popper (1972, 1994) who argues the importance of formulating things in written language for a critical view of knowledge. For Popper (e.g. 1972: 73), 'objective knowledge' includes theories which are published in journals and books, the logical content of books, libraries, computer memories, and so on. Swales provides substantial textual and ethnographic evidence for the phenomena which Popper discusses. (See chapter 9.1.)

Concepts which seem underdeveloped in Swales' study are

- the mediation of knowledge by language
- the comparative analysis of a text type against the
- background of variation in the language as a whole.

These three studies are very different in style, and have different strengths and weaknesses. Other important case studies are discussed elsewhere: see chapter 5, Appendix, for a discussion of Berk-Seligson's (1990) analysis of a corpus of courtroom data.

1.8. SUMMARY

As a group, these three studies raise a range of questions which have to be tackled by a theory of text interpretation:

- the relations of text to corpus, and of instance to system
- the concepts of sample, representativeness and typicality
- the probabilistic nature of many textual patterns
- the meaning of patterns of frequency and distribution
- the function of syntactic units in texts
- the mediation of knowledge in language
- the dissemination of texts from authors to audience.

I will discuss all of these questions below, and propose techniques for making some progress with them.

So, this book is about the computer-assisted analysis of text semantics. It represents part of a research programme on the relations between texts, text types, text corpora and social institutions. The linguistic focus is on lexical and grammatical patterns in texts, particularly those patterns which express the point of view of the speaker or writer. The descriptive focus is on texts which are public and/or authoritative, such as newspaper articles, school textbooks, government reports and legal judgements. This leads to a sociological focus on the relation between the micro structure of texts and the macro structure of social institutions, such as schools and courtrooms. Texts, spoken and written, comprise much of the empirical foundation of society: they help to construct social reality. And textual analysis is a perspective from which to observe society: it makes ideological structures tangible.

Much of this deep patterning is beyond human observation and memory. It is observable only indirectly in the probabilities associated with lexical and grammatical choices across long texts and corpora. This therefore leads in turn to a methodological focus on computer-assisted and quantitative methods, particularly in cases where native speaker intuitions are very limited, and where description can proceed only on the basis of attested corpus data.

The major intellectual puzzle in the social sciences is the relation between micro and macro. How is it that routine everyday behaviour, from moment to moment, can create and maintain social institutions over long periods of time? (See chapter 3.4, 3.5.) Institutions are networks of commitments: they are based on contracts, promises, requests, authorizations and other speech acts. And the many things created by texts and by ways of talking include the mass media, education, government, and the law. Text and corpus analysis provide methods for studying the empirical basis of society in these micro-macro relations.

The rest of the book now presents some history and principles of text and corpus analysis (chapters 2 and 3) and some computer-assisted analyses of texts and corpora (chapters 4 to 8).