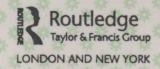# The Linguistics of Political Argument

The spin-doctor and the wolf-pack at the White House

Alan Partington

The Linguistics of Political Argument

The spin-doctor and the wolf-pack at the White House

How does the White House 'sell' its message to the press? How hostile or docile are political reporters? Do they really try to 'transcend the Spin to find The Truth' (*Business Week*), or do they have agendas of their own? How many roles does the spokesperson (or 'podium') enjoy and how many the journalist? How are political – and personal – arguments fought and resolved? What rules of politeness prevail? How does the podium employ 'invisible' metaphors to constrain the audience's world view? Is mode   precedent in standard rhetoric?

This book examines the relationship between the White House and the person of its press secretary, and the press corps through an analysis of the language used by both sides. A corpus was compiled of around 50 press briefings from the late Clinton years. A wide range of topics are discussed from the Kosovo crisis to the Clinton–Lewinsky affair. This was a particularly intriguing and stressful time in the White House–Press relationship.

The work is highly original in demonstrating how concordance technology and the detailed linguistic evidence available in corpora can be used to study discourse features of text and the communicative strategies of speakers. It will be of vital interest to all linguists interested in corpora, discourse and pragmatics, as well as sociolinguists and students of communications, politics and the media.

Alan Partington is Associate Professor of Linguistics in the Faculty of Political Science, Camerino University (Italy). He has published in the fields of phonetics, CALL, lexicology and corpus linguistics, and is the author of *Patterns and Meanings: Using Corpora for English Language Research and Teaching* (1998). He is currently researching ways in which corpus techniques can be used to study features of discourse.

Routledge Advances in Corpus Linguistics
Edited by Anthony McEnery
Lancaster University, UK. and
Michael Hoey
Liverpool University, UK

Corpus based linguistics is a dynamic area of linguistic research. The series aims to reflect the diversity of approaches to the subject, and thus to provide a forum for debate and detailed discussion of the various ways of building, exploiting and theorizing about the use of corpora in language studies.

The Linguistics of Political Argument

The spin-doctor and the wolf-pack at the White House

Alan Partington

© 2003 Alan Partington

Contents

Foreword

The spin-doctor and the wolf-pack

The players, the peculiarities and the political importance of White House press briefings

A quite remarkable variety of metaphors have been employed by commentators, many of whom are unsympathetic towards the participants, in describing the briefings held daily at the Office of the White House Press Secretary. They are 'a political chess game' (Reaves White), in which 'both sides view everything the other side does as a mere tactic' (Kamiya). Alternatively, they are 'rhetorical combat' (Kurtz), a 'war zone' in which 'combatants with a multitude of agendas [...] prepared for battle' (Reaves White). They are 'a wrestling match' and a duel or 'face-off' (Reaves White) but also 'a weird formulaic dance' (Kamiya).[1]

The spokesperson (or podium)[2] is a soldier under 'hostile media fire [...] on the front lines for Clinton on nearly every major battle (Baker and Kurtz) but also a sailor who must 'navigate the treacherous waters of the daily briefings' (CNN-allpolitics) and is frequently found 'desperately scrambling and bailing to keep a torrent of scandals from sinking the battered ship of state' (Jurkowitz). He[3] is both a pugilist who has 'bobbed and weaved and jabbed [...] his way through all manner of Clinton scandals' but also a street thug who 'beats up on reporters' (Kurtz). He has even tried to be 'an ambassador between a president who disdained the press and reporters who didn't much trust the president' (Kurtz). Less nobly, he is 'a propagandist and a smear artist' (Irvine and Kincaid), 'a master at keeping the press in its place by doling out exclusives to reporters who will play ball with him and actually sandbagging those who will not' (Zweifel). Above all, he is a 'spinmeister extraordinaire' (Kurtz) eternally spinning the truth, whatever that might be.

There are slightly fewer metaphors to describe the journalists, probably because most of the commentators are themselves press people. Nevertheless, they are wild animals, the 'wolf-pack'[4] of my title, which 'fights over morsels' (Warren). They too can be boxers out to 'pummel' the spokesperson who has 'to stand at the podium and take whatever abuse the fourth estate wanted to dish out' (Kurtz). They are 'cynical chroniclers' (Kurtz), 'petty and manipulative [who] simply cannot put aside their "gotcha" mentality' (Zweifel). 'They like to destroy people. That's how they get their rocks off' (Dunham, supposedly quoting President Clinton). At the same time, however, they are 'a lot of dupes' (Irvine and Kincaid)

and 'the White House reporter is not much more than a well-compensated stenographer' (Warren). On a more exalted note, their 'job is to transcend The Spin to find The Truth' (Dunham).

Hopefully, we will be able to judge the validity of these metaphors in the course of this book. What is beyond dispute is that these briefings comprise a particularly fascinating genre of institutional talk in which the two parties involved, the podium and the press, have very different interests and aims in life, which are in conflict on several levels. The podium wishes to project his political ideas and particular view of the world, the press to test that view – to destruction if necessary. The press hopes to uncover information, any evidence of weakness, malpractice, internal dissension and so on, the podium to give as little as possible away outside the official line. They adopt and exploit different participant roles (or *footings*, Chapters 2–4), command non-symmetrical sets of discursive resources, and employ different discourse strategies (Chapter 8), they use different metaphors to describe the world (Chapter 10) and probably even see the whole nature of the business being conducted in different ways (Section 7.8). These differences become so severe at one point – during the Clinton–Ms Lewinsky affair and the near impeachment of the President – that, as we shall see, communication between the two sides comes close to breaking down.

Despite these briefings being an instance of *conflict talk* (Chapter 8), however, and notwithstanding a recognition on both sides of their adversarial nature (Chapter 6), the protagonists know each other so well, and probably *need* each other so much,[5] that the register of the briefings is generally highly familiar. This state of affairs leads to particularly complex and intricate relations of politeness (in the Brown and Levinson sense of getting the best out of social relations) (Chapter 7).

What transpires in these briefings can also be extremely important and highly delicate from a political perspective:

> Anything McCurry (press secretary at the beginning of the Clinton administration) uttered from the podium magically attained the status of official White House policy, and if he deviated later on, the administration would be accused of the dreaded sin of flip-flopping.
>
> (Reaves White)

Not only are the podium's words often treated by the press as White House policy, but they risk interpretation by non-American bodies as official *US* policy. And since they are broadcast both on television and on the Internet, 'any misstep can be beamed instantaneously around the world' (CNN-allpolitics). A word out of place, a mistimed remark, even a simple oversight can have disastrous political or diplomatic repercussions. All this exposure, of course, means fame: 'the chief White House spokesman's face is probably as well known as any cabinet member' (CNN-allpolitics). In Galtung and Ruge's (1981) terms he is 'newsworthy', has become an 'élite person' in his own right. Many of the journalists, too, are well-known television faces or newspaper by-lines.

Acknowledgements

0 Introduction

Corpora, discourse, politics and the press

0.0 The three topics of this book

This book has three topics. It is principally a study in *corpus linguistics*, in particular, of how features of spoken discourse, including rhetorical strategies, can be analysed with the aid of corpora. It is also a study of a certain type of *institutional talk*, that of press briefings. Third, it is a case study in the relationship between *politicians and the press*. Here, we will discuss the first and last of these themes; the second will be treated in Chapter 1.

0.1 The corpus of briefings

0.1.1 Source

The research described in this work was conducted on a corpus of 48 briefings, comprising approximately 250,000 words of running text, whose transcripts were downloaded from the White House Library website, where they are openly available for inspection. The vast majority date from the period between 1996 and June 1999, that is, before and during the Kosovo crisis (although many other topics are discussed, from the budget to the Clinton–Ms Lewinsky case). This was a particularly intriguing and stressful time in the relationship between the White House and the press. Many White House briefings are also transmitted over the Internet by C-Span, a US public-service broadcaster. I was able, therefore, to watch a number of them and to check the nature and quality of the transcripts.[1]

0.1.2 Transcription

The method of transcription used is fairly broad, there being no phonetic detail. The transcriber attempts to reproduce the normal punctuation of written English texts. They are further idealized in the following ways. Brief repetitions or recapitulations are omitted so that, for example, 'I beg your pardon, beg your pardon ...' becomes simply 'I beg your pardon'. Intercalations such as *y'know,*

er, em are all missing, as are some (but not all) replannings, for example, 'the US has the right to make to suggest...' becomes 'the US has the right to suggest...'. This has the unfortunate result that moments of hesitation, of tentativeness, tend to be hidden from the transcript reader. Inaudible segments are marked with long hyphens but there is no indication of their duration. When two press voices are competing for a question turn, the transcriber, when possible, includes both, one below the other. When a press voice coincides with the podium's, the latter is usually the one that gets written, largely because his is clearer – he is the one with the microphone. Laughter is indicated by [Laughter] at the end of the turn. There is no indication of voice quality, for example, regional or foreign accents, loudness–softness, speed. Although the linguist may bemoan many of these transcription choices, the intention has clearly been to make the written texts as easy as possible to read. As Cook (1995: 45) remarks, the degree of detail in transcription is analagous to mapmaking, a question of choice in regard to the need of the user. The transcripts also appear to be complete in the sense that what is broadcast is transcribed. There is no editing out of compromising or embarrassing episodes. Finally, the transcriptions are produced and published on the White House website very quickly. Occasionally, what appear to be typing errors occur, but I have made no attempt to correct what I found. I have, however, often edited examples for reasons of space.

0.1.3 Research tools

A number of semi-automatic computational tools were used to help analyse the corpus. The most important of these is undoubtedly the *concordancer*. This is essentially a collector and collator of examples. It will search a text or set of texts for a string of letters (keyword or phrase) and present all examples in a list with a certain amount of co-text for each one. Such a list enables the analyst to look for eventual patterns in the surrounding co-text, which may provide information about the use of the key-item. These lists can be prepared and edited in several ways. The entries can be *sorted*, that is, listed in alphabetical order, according perhaps to the first word to the left (which, if the keyword is a noun would group together the adjectives preceding it), or to the right (which, if the keyword is an adverb would group the following adjectives). Unwanted lines can be removed by a *cancel* (or *zap*) facility. Most of the concordances presented in this work are *sentence concordances*, which are often easier to read than a 'crude' concordance. The lines have been saved into a text file and subsequently manually edited so that the co-text for each key-item is more or less the whole transcript sentence it appears in.

The concordances reported in this book were prepared using either *MicroConcord* (Scott and Johns 1993) or *WordSmith Tools* (Scott 1998). This latter also provides a program to prepare frequency lists; that is, the words in a corpus can be listed in order of how frequent they are in the corpus. The following are

the 30 most frequent word *types* in the briefings corpus:

N	Word	Freq. (%)	N	Word	Freq. (%)
1	THE	15.174	16	HAVE	2.289
2	TO	8.603	17	LOCKHART	2.093
3	THAT	7.400	18	IT	2.049
4	OF	5.924	19	BE	1.841
5	AND	5.813	20	FOR	1.806
6	A	4.328	21	ARE	1.745
7	IN	4.078	22	HE	1.731
8	IS	3.471	23	PRESIDENT	1.731
9	I	3.466	24	NOT	1.661
10	MR	3.454	25	WHAT	1.649
11	Q	3.399	26	THINK	1.608
12	WE	2.893	27	WITH	1.600
13	YOU	2.578	28	THERE	1.562
14	ON	2.534	29	AS	1.557
15	THIS	2.328	30	WILL	1.514

Thus, just over 15 per cent of all word *tokens* in the corpus are the word type *the*, and so on. Note the huge preponderance of grammatical words. This is the case in any corpus though the precise order of the words, their places in the table, can vary.

The frequency lists of two or more corpora can also be compared using the *Keyword* facility to show up *relative* frequency, or *key*-ness of vocabulary in a corpus; it should be noted that this is a different use of *key* from that used in concordancing. This process will be explained in more detail in Section 0.4.

A number of other corpora were utilized in the course of this research as a basis for comparison with briefings discourse. These include several corpora of journalistic texts: a collection of British news interviews (*INTS*) of similar size to the briefings corpus, a four-million word corpus of written newspaper texts (*Newspaper Corpus*), the CD-ROM of all articles in the *Times* from 1993 (*Times*). The *Brown*, *Frown* and *Flob* corpora and the Wellington spoken corpus (*WSC*) of general conversation were all used when appropriate.[2] The *British National Corpus* (*BNC*) on the Web was also occasionally consulted.

0.2 Corpus linguistics and discourse

This book is proposed in the first place as a contribution to the field of corpus linguistics. It attempts to show how it is possible to use concordance technology and the detailed linguistic evidence available in corpora to enhance the study of the discourse features of a particular genre of the language. It is especially an investigation of the communicative strategies used by speakers to pursue their designs.

0.2.1   *Studying discourse using corpora*

This makes it an unusual contribution to the discipline. In comparison with the impressive strides corpus linguistics has made in the fields of lexicography, grammatical description and register studies among others (see McEnery and Wilson 1996: 87–116; Kennedy 1998: 88–199; Partington 2001: 46–63), it has had relatively little to say in describing features of discourse, particularly of inter-action, that is, the rhetorical aspects of texts. Much research on (large, hetero-generic) corpora strives to make observations about 'the language' with the goal of explaining and improving human understanding of it, and has practical proj-ects in mind, such as building more accurate grammars or of generating more informative dictionaries. Language in these endeavours is treated as *product*, as a preexisting entity, deliberately and of necessity removed from its context of production, from the *processes* of its authorship and reception. The aim is to get to know the product as well as possible.

On the other hand, 'the amount of corpus-based research in pragmatics and discourse analysis has to date been relatively small' (McEnery and Wilson 1996: 98). Biber *et al.* (1998) also note that discourse studies 'are not typically corpus-based investigations'. There are a number of reasons for this. 'Pragmatics and discourse analysis rely on context – pragmatics has often been defined as "mean-ing in context" – whereas corpora strip much of the context of utterances' (McEnery and Wilson 1996: 98). Many corpora have also tended to take frag-ments of texts rather than whole texts, whereas the latter are necessary for many types of discourse analysis. There has until relatively recently been a paucity of spoken corpora to use for spoken discourse analysis.[3] Although some modern corpora like the *BNC* include socially relevant information (speaker's sex, age, geographical extraction, etc.), it is still up to the analyst to relate systematically such information to what speakers are attempting *to do* at any moment in their interaction.

The present work however, attempts to study discourse strategies by reintro-ducing as much attention to context as possible – including both the wider historical, political, mediatic context, and the more immediate local and personal contexts, especially the participants involved, the speakers and hearers. Moreover, attention to the *co-text* is maximized by having whole texts, whole briefings acces-sible in both mechanical and non-mechanical form, so that they are available for various different methods of analysis. It attempts, thus, to 'take into account both product and process: not only the text itself, but also its production and reception' (Stubbs 1996: 8). Such an approach is feasible with specialized corpora, corpora of a single text-type (monogeneric), where these processes and the contexts they take place in remain relatively constant, or at least alter in relatively predictable ways (see the discussion of specialized corpora in Chapter 13).

Questions of authorship and reception are, therefore, of central interest in this book. So too are the features of *interaction* between the 'online producers/ receivers', that is, the speakers: the strategies they adopt, the aims they have and the effect they produce on each other. The main aim of this kind of corpus

research is no longer to investigate the contents of the corpus as an objective in itself but as a means, an instrument – along perhaps with others – to study what discourse participants are doing in particular circumstances. Specialized corpora are also increasingly used in this instrumental way in the field of education (Flowerdew 1993; Granger 1993), including translator training (Zanettin 1994). Corpora of scientific English, business English, computer science and many other text-types have been compiled in ESP contexts in order to teach them better (McEnery and Wilson 2000: 201–8).

In addition, this work also recognizes the need to study discourse production in its social and political context, that is, to investigate the relationship between texts, authors and the social institutions which produce them, of which they are a part and which they help to reproduce (i.e. produce anew, Fairclough 1989: 39–40). Since social institutions are always defined and supported by particular texts in this way, there is a natural link between institutional analysis and textual analysis. Nowhere is this relationship more clear than in these texts – in the briefing, the discourse *is* the institution, the use of language is the whole point of the enter-prise. This book then is also meant as a contribution to the study of institutional talk, of which more later.

The present work, thus, has two main areas of concern. The first, the linguistic, 'is on lexical and grammatical patterns in texts, particularly those patterns which express the point of view of the speaker' (Stubbs 1996: 20). Beyond this, I add those patterns that reveal relationships between speakers, their strategies of persuasion, the metaphors of the world they adopt, and the divergencies between what a speaker professes and what they really seem to think. The problem remains, however, of elaborating a suitable methodology to study features of interaction in large bodies of texts, in corpora. Or put in another way, of designing ways of using corpus-analytical techniques to study aspects of discourse. Much of this book will be an attempt to devise such methodologies.

The second, the institutional concern, is the analysis of language in texts that are public and/or authoritative, that is, the emanation of institutions of power and influence. The particular institutions in this study are the government and the press, the first and fourth Estates, and White House briefings are public and authoritative in the ways outlined at the end of the Foreword. *Authoritative* can also have another sense, that of 'convincing'. An important question in any study of institutional language is 'how is discourse organized to appear factual, literal, objective, authoritative and independent of the author, rather than appearing to be expressed from a particular point of view?' (Stubbs 1996: 97). As we will see, the podiums in these briefings have to make an extraordinary effort to organize their discourse to these ends, since the 'authoritative', official point of view is constantly under attack. The process by which institutions attempt to construct, reproduce and constrain social reality is neither wholly deterministic nor conducted by a single power group. It has to be negotiated between parties and this negotiation is often highly conflictual. These briefings are an excellent

example. In my view, they are worth studying precisely because of the interplay between the institutional aims of the two Estates and the discursive strategies adopted by the individuals involved to achieve these aims, together with the fact that the conflictual nature of proceedings requires participants to constantly refine their strategies of persuasion and resistance to persuasion, of evasion and pursuit.

0.2.2 Why utilize a corpus?

The general value of corpora has been discussed at length (McEnery and Wilson 1996; Kennedy 1998; Partington 1998) and can be summarized as follows:

> text corpora provide large databases of naturally occurring discourse, enabling empirical analyses of the actual patterns of use in a language, and, when coupled with (semi-) automatic computational tools, the corpus-based approach enables analyses of a scope not otherwise feasible.
>
> (Biber *et al.* 1994: 169)

Examples of authentic data can serve to support the researcher's argument or, perhaps even more importantly, as counter-evidence to make them think again. The use of corpora has profound philosophical implications for the kind of critical analysis of discourse we are engaged in here, as Hardt-Mautner argues: 'drawing on corpus evidence fundamentally redefines the nature of "interpretation"' (Hardt-Mauther 1995: 22), because it adds an empirical dimension to introspection and marries quantitative and qualitative research methodologies (Haarman *et al.* 2002).

Checking intuition against naturally occurring instances of language also frequently serves as a springboard for new intuitions, and can open up new and previously unexpected avenues of thought. This occurs because much of what carries meaning in texts is not open to direct observation: 'you cannot understand the world just by looking at it' [Stubbs (after Gellner 1959) 1996: 92]. One of the reasons why this occurs is that language is used semi-automatically. It is well known that even authors themselves are seldom fully aware of the meanings their texts convey (and not just verbal ones; the same is true of the visual and the musical). This is because meaning is spread out in all parts of a text, in far more tiny parcels than is usually realized. It is conveyed in all the choices the author makes at the lexical level – for example, whether to use *liberty* or *freedom* or the more religious *deliverance*. But it is conveyed in all the choices at the grammatical level too: whether to express a process as a verb or a noun; a description as an adjective or an adverb; what roles to give participants in an event [are they perhaps actor or reactor?; are they grammatical subject and/or thematic (or psychological) subject?]; when an utterance contains two ideas, whether to present them as coordinate (paratactically) or as one subordinate to

the other (hypotactically); which order to present them in (i.e. which to *thematize*); how much and what kind of modality to employ, that is, to state how certain or frequent or necessary something is, or how able, willing or obliged someone is to do it. And these choices can be mutually supporting (such as when an entity is *always* made actor), or can be independent (random) or can indeed be contrasting [e.g. an entity is first of all made actor, then switches to acted on (perhaps *victim*) half way through]. In the first and last cases, of course, since the choices are *patterned*, their meaning becomes denser and more significant. However, the text receivers may be unaware that anything is going on at all. Section 0.2.3 shows how automatic text analysis tools can throw into relief the non-obvious in a single text, can shed light on what may be hidden thoughts, hidden perhaps even from the author.

0.2.3 Grammar and hidden thoughts: concordancing the Declaration of Independence

The way in which concordance technology can be used to examine a text to highlight how its authors see or project the world is illustrated in the following brief analysis of the Declaration of Independence of the United States.

Of all the grammatical choices an author makes, perhaps the most consistently important are those regarding *transitivity*. 'Transitivity is the foundation of representation: it is the way the clause is used to analyse events and situations' (Fowler 1991: 71). It is a consequence of the linear/temporal nature of language. It both *forces* authors to structure in linear fashion what they might originally have perceived as a unitary situation and also *allows* them to portray reality in a way they might wish, to construct an argument in syntax.

A number of authors suggest that a fruitful way to begin the study of transitivity in a text (i.e. to find out who is doing what to whom) is to look at the use of pronouns. The author(s) of the Declaration (mainly Thomas Jefferson)[4] claim that their argument is with the King of Great Britain (rather than the British people), and all references to *he* in the document refer to the King. The concordance of the word *he* is as follows:

1	egislate for us in all cases whatsoever. He has abdicated Government here, by declaring
2	ithout the Consent of our legislatures. He has affected to render the military independ
3	them and formidable to tyrants only. He has called together legislative bodies at plac
4	nt of and superior to the Civil Power. He has combined with others to subject us to a
5	worthy the Head of a civilized nation. He has constrained our fellow Citizen taken Ca
6	m into compliance with his measures. He has dissolved Representative Houses repeate
7	rom without, and convulsions within. He has endeavored to prevent the population of
8	amount and payment of their salaries. He has erected a multitude of New Offices, and
9	, or to fall themselves by their Hands. He has excited domestic insurrections amongst
10	e and necessary for the public good. He has forbidden his Governors to pass laws of
11	ur people, and eat out their substance. He has kept among us, in times of peace, Standi
12	ws for establishing Judiciary Powers. He has made Judges dependent on his Will alon

13 tions of new Appropriations of Lands. **He** has obstructed the Administration of Justice
14 rotection and waging War against us. **He** has plundered our seas, ravaged our Coasts,
15 s utterly neglected to attend to them. **He** has refused to pass other Laws for the accom
16 Facts be submitted to a candid world. **He** has refused his Assent to Laws, the most wh
17 s invasions on the rights of the people. **He** has refused for a long time, after such dissol
18 , and destroyed the lives of our people. **He** is at this time transporting large armies of for

There are no occurrences at all of the word *him* in the document. It is immediately clear that *he* is always sentence initial, and, of course, is indicative of subject position in the phrase. The King of Britain is, therefore, presented invariably as the active protagonist of the situation. The concordance is sorted according to the second word to the right of the keyword in order to highlight the verb following the *he* subject and *has* auxiliary. As can be seen they are generally unpleasant in connotation (*abdicated, constrained, forbidden, refused, obstructed*, etc.), occasionally violent (*excited ... insurrections, plundered*).

The authors refer to themselves, or their supporters,[5] using *we*.

1 be the ruler of a free people. Nor have **we** been wanting in attention to our British brethren.
2 ir native justice and magnanimity, and **we** have conjured them by the ties of our common k
3 s. In every stage of these Oppressions **we** have Petitioned for Redress in the most humble
4 ng in attention to our British brethren. **we** have warned them from time to time of attempts
5 an unwarrantable jurisdiction over us. **We** have reminded them of the circumstances of our
6 of our emigration and settlement here. **We** have appealed to their native justice and magnan
7 ces our Separation, and hold them, as **we** hold the rest of mankind, Enemies in War, in Pea
8 s which impel them to the separation. **We** hold these truths to be self-evident, that all men
9 voice of justice and of consanguinity. **We** must, therefore, acquiesce in the necessity, whic
10 the Protection of Divine Providence, **we** mutually pledge to each other our Lives, our Fort
11 , Enemies in War, in Peace Friends. **We**, therefore, the Representatives of the United Stat

Many of the verbs following *we* are verbs from the semantic fields of warning and reminding, for example, *appealed, petitioned, reminded and warned*. *We* is projected as a thoroughly responsible group, if a little schoolmasterly. There are an equal number of occurrences of *us* (11) where the *us* is the object of the action, the doing of which is performed by the *he*. There is a consistent contraposition of these two entities, the one who does and the other who is done to. This, of course, is consonant with the view of the world the authors wish to propound of a despotic King mistreating a blameless population.

However, by far the most common form is the possessive *our*.

1 Nor have We been wanting in attention to **our** British brethren. We have warned them from
2   For taking away **our** Charters, abolishing our most valuable Laws
3 grid0 He has plundered our seas, ravaged **our** Coasts, burnt our towns, and destroyed the li
4 , and we have conjured them by the ties of **our** common kindred to disavow these usurpatio
5 rpations, which would inevitably interrupt **our** connection an d correspondence. They too h

6   For imposing taxes on us without **our** Consent:
7 ers to subject us to a jurisdiction foreign to **our** constitution, and unacknowledged by our la
8 ve reminded them of the circumstances of **our** emigration and settlement here. We have ap
9   He has constrained **our** fellow Citizen taken Captive on the high Sea
10 mutually pledge to each other our Lives, **our** Fortunes and our sacred Honor.
11 endeavored to bring on the inhabitants of **our** frontiers, the merciless Indian Savages, wh
12 and altering fundamentally, the Forms of **our** Governments:
13 me Judge of the world for the rectitude of **our** intentions, do, in the name, and by authorit
14 our constitution, and unacknowledged by **our** laws; giving his Assent to their acts of prete
15 , Standing Armies without the Consent of **our** legislatures.
16 idence, we mutually pledge to each other **our** Lives, our Fortunes and our sacred Honor.}
17 For taking away our Charters, abolishing **our** most valuable Laws, and altering fundamen
18   For suspending **our** own Legislatures, and declaring themselves
19 sent hither swarms of Officers to harass **our** people, and eat out their substance.
20 urnt our towns, and destroyed the lives of **our** people.
21 ed for Redress in the most humble terms: **Our** repeated Petitions have been answered onl
22 to each other our Lives, our Fortunes and **our** sacred Honor.
23   He has plundered **our** seas, ravaged our Coasts, burnt our towns,
24 uiesce in the necessity, which denounces **our** Separation, and hold them, as we hold the r
25 dered our seas, ravaged our Coasts, burnt **our** towns, and destroyed the lives of our peop
26   For cutting off **our** Trade with all parts of the world:

In every other corpus I have examined, *our* is only half as frequent as *we* or even less. Their proportions here are, therefore, highly marked. Once again, the author of the action and the grammatical subject is generally *he*. Among the entities affected by the action are the rights (*our laws, our Governments, our constitution*, etc.) and the possessions (*our towns, our Fortunes, our Trade*, etc.) of the group referred to with *we*. Thus, the authors of the document achieve two ends. They emphasize that *we* are already a separate group from the British; that *we* enjoy, as the document itself proclaims, a 'separate and equal station to which the Laws of Nature and Nature's God entitle them'. (On only two occasions – lines 4 and 5 – is *our* inclusive referring to the British together with the *we* group. On the other 24 occasions, it is exclusive and refers only to the latter.) They also reveal, however, the degree to which their dispute with the British government is about property, as much about claiming *our* rightful possessions as about political rights.

How consciously Jefferson made all these transitivity choices is a matter for conjecture. It seems likely that the active portrayal of the King was deliberate. Whether or not he also meant to portray his companions' motivation as quite so business-like is less certain (although eighteenth century liberal thought was much less coy than it is today about the connection between property and freedom). What we can say for sure is that concordancing can reveal patterns even within a single text that throw light on its meaning and function, and also on the conscious and unconscious strategies adopted by authors. A detailed examination of the use of pronouns in the White House briefings is contained in Chapters 3 and 4.

0.2.4 Comparing instance with system, comparing genres

We began this discussion with the observation that much of what carries meaning in texts is not open to immediate observation. If this is true of a single short text then it is true *a fortiori* of a collection of texts, or of the whole genre it belongs to, or of the collection of genres known as a 'language'. Human observation skills, attention-span and memory just cannot process the mass of data directly. And, of particular importance to the present research, only by comparing the choices being made by speakers at any point in a text with those which are normal, usual within the genre can we discover how *meaningful* those choices are. If an author/speaker does something that is normal, this carries little meaning. They have simply followed the normal dictates of the local style. But if, as we have seen in the analysis above, they do something unusual (like making one of the protagonists unvaryingly phrase subject), a special choice has been made and the analyst must search for the reason. To be able to compare the patterns found in a single text with those usual in the genre, to compare, in Halliday's terms (1992), *instance* with *system*, a corpus of texts which, at least to some small degree, *represents*[6] the genre is invaluable. Furthermore, if we are also, as here, interested in the characteristics of the genre itself, it is vital to be able to compare its particular features and patterns with those of other genres. In this way we discover *how* it is special, and can go on to consider *why*. All genre analysis is thus properly comparative. If texts are not compared to other bodies or corpora of texts it is not possible to know or to prove what is normal. Only against a known background of what is normal and expected can we detect the unusual. And, as we said, it is the unusual and unexpected which carries meaning.

0.2.5 Corpora and context

To investigate the precise meaning of a single linguistic event, then, we need to be able to study it in its context. If, moreover, we are interested in participants' strategies of interaction, we need to know a great deal about the context of that interaction. We return, for a moment, to the discussion of context of production and reception.

A number of authors, Baldry in particular, have criticized corpus linguistics in general and the concordance, its main tool of analysis, in particular, as abstracting text from its context, of failing to include 'any criteria for showing how language, in context, systematically integrates with other semiotic resources. In this sense corpus linguistics continues to treat language as a self-contained object' (Baldry 2000: 36).

The question is a complex one and involves two related phenomenological issues. The first is that all transferral of data from one medium to another will entail loss. One need only think of the reductiveness of trying to describe music in words. A videotape of a wedding is not a wedding. The second is that all description implies abstraction (the only non-abstract description of the

world is the world itself) and all abstraction also means loss of detail. The greater the abstraction (i.e. the higher the level of overview), the more of the object of study can be included but the more individual contextual detail will be lost. Conversely, lowering the level of abstraction increases the contextual detail but reduces the portion of the world that can be included for the same degree of time and effort spent on the compilation. Moreover, from an analyst's point of view, too much detail means no overview, no possibility of generalization. To revive Cook's (1995) map-making metaphor, a map of the world that depicted everything in the world would have to cover the whole world. And would be of no use to anyone.

The researcher must first of all choose the level of abstraction that best serves their purposes. If we need to find our way from the Colosseum to the Trevi fountain a map of Europe is useless. Conversely, a street plan of Italian cities will tell us precious little about European national borders. To understand a podium's in-joke, we must look at a single text in all its contextual detail, the personalities involved, what has just occurred and so on. But to design a grammar of register (such as Biber *et al.* 1999), we must exploit large heterogeneric corpora and quantitive analysis.

This said, the most desirable system of all is one that offers the greatest possible flexibility of research methods and mobility between analytical levels. Baldry (2000: 30) argues that it is important to ascertain 'what linguistic corpora have left out of the picture, and how might these other factors be re-instated in the processes of data collection, transcription and analysis'. In other words, he argues that this flexibility should be built into the design of corpus compilation from the very beginning to make it feasible and simple to move between levels of analysis. For spoken texts, Baldry and his team are in the process of devising a means of transcribing some of the visual and kinetic aspects of production: movement, gesture, gaze, colour and so on.

The methodology I adopted for this research attempted, as far as was possible for me at the time, to cater for this need to move between levels of analytic detail. The corpus comprises separate files, each containing a briefing coded by date and initials of briefer. This enables easy movement from concordance line (where file names are indicated) to the individual file. A higher level of abstraction even than the concordance are the word frequency lists prepared using *WordSmith Tools* for both the briefings corpus and a number of other corpora I decided to use as comparisons (Section 0.4). At the lower levels, I was not able, at the time of compiling, to download videos of the briefings. However, I was able to watch recordings broadcast through *Realplayer* and make notes. Researchers who might wish to replicate this work on similar data can still follow the briefings on C-Span. Nevertheless, not having the recordings means I cannot, if in doubt, check back on voice quality, intonation and so on – a regrettable loss. Increasing computing speed and memory capacity is now beginning to make it possible for the analyst of a spoken corpus to move from concordance to transcript to audio or audio-visual recording of the speech event.

0.2.6 Combining methods of analysis

One highly unusual step in my methodology was that I spent a summer reading the corpus of briefings and making notes that were transferred onto disk and were thus themselves available for concordancing. This made it quick and simple to collect at a later moment all instances of any phenomena I had noted, say, oddities of collocation or instances of question avoidance. The size of the corpus was set deliberately at a quarter of a million words in order to be large enough for meaningful patterns to emerge and small enough for a sense of it to be acquired by being read in this way. These notes clearly constitute an intervention on the part of the researcher into the object of research, but this is inevitably present at all levels (including, we should not forget, concordance preparation and interpretation).

One of the objects of research was, in fact, to compare the kinds of information that can be derived from reading a set of texts with the kinds that derive from semi-automatic analyses, to determine what can be done better with 'non-automatic' procedures and what is better performed with computer assistance. In most cases, it turned out to be a question of the one reinforcing rather than replacing the other, of results being obtained by a heuristic combination of the two. Examples of this will be found throughout the book. A few immediate examples are the following. The prereading alerted me to the phenomenon of other-reformulation, but only the concordance revealed its extent and complexity, the mass of phraseology used to perform it (Chapter 9). In Chapter 7, concordancing the same modal expression (e.g. *just*) separately in the podium moves and the press moves highlights significantly different patterns of use. I then had to go on to hypothesize why – what different strategies these patterns might serve and for this it was necessary to look at the text transcripts in detail. In Chapter 10, comparing frequency word lists derived from different corpora sheds much light on the special metaphors used in briefings. Finally, I noticed during the prereading an odd use of the word *adamant*: 'President Milosevic has proved to be quite *adamant* in his campaign of atrocities'. Concordancing the word in larger corpora revealed why I felt the use was unusual (Section 0.3.3). But why the podium used it remains a matter of speculation.

Thus, at the simplest level, corpus technology helps find other examples of a phenomenon one has already noted. At the other extreme, it reveals patterns of use previously unthought of. In between, it can reinforce, refute or revise a researcher's intuition and show them why and how much their suspicions were grounded.

0.3 Concordancing for indicators of point of view

Many writers have noted that there are a number of linguistic indicators that are particularly useful in revealing an author's opinions, attitudes and ideology. Most of them draw on the ideas of Halliday (1973, 1985), who conceives of grammar

itself as a 'social semiotic'. The very categories that natural language grammars encompass have arisen as a response to the need of human animals to express certain meanings to each other (i.e. socially) through the linear medium of language. The categories of language are thus *functional* in this sense.

We have already seen an example (see Section 0.2.3) of how the transitivity system can be used to express points of view, and how the concordance can, given the right circumstances, be exploited to reveal some of them. Other linguistic areas that can often usefully be studied are those of modality, the numerous vagueness or information-concealment techniques and semantic prosody.

0.3.1 Modality

It is a relatively simple matter to concordance modal operators such as *can, may, might, could, would, should* and so on. A certain amount of editing will be necessary to remove unwanted examples (such as when *May* refers to the month). It is beyond the scope of this book to launch into a full-scale analysis of modality in discourse. But one or two interesting statistical phenomena were noticed regarding these briefings

First of all, I prepared two subcorpora by separating the questioner moves from the podium response moves contained in four of the briefings (around two hours of talk) and placing them in two separate files, the Q-file and the R-file, respectively. The importance of comparison in linguistics has already been stressed and this division made it possible to compare and contrast press and podium language. These files were used for several studies in this book.

In the Q-file, the word *might* was found to be much more frequent than *may* (nine occurrences to one); in the R-file, however, the reverse was the case, *might* was far less frequent than *may* (two to eleven). If one ignores examples of the questioner's typical permission-seeking *if I may follow up on that*, these proportions seem to be regular throughout the corpus – *may* being far more frequent in the responses, *might* in the questions. The explanation for this could be that the slightly more distancing effect of *might* makes it generally more suitable for questioning, which is notoriously face-threatening (Chapter 7) for the respondent. The preponderance of *may* over *might* in the kind of defensive responses typical of these briefings could, on the other hand, express a desire to reflect control over the situation – 'we may do so if we choose' – *we* being the predominant subject of *may* in podium replies.

By using the item *Q* as a context word, it was possible to concordance modals in the first few words (set randomly at 10) of the question moves. One interesting finding was that, of the *necessity* modals, *should, have to* and *need* are found in this corpus to ask questions, whilst *must* and *ought to* are not.[7] All five items were common in the podium's moves, probably because they are typical of opinion-giving:

(1) MR McCURRY: And we think the benefits of tax relief *ought* to be focused on low and middle income

(2) MR LOCKHART: [...] Furthermore, Israelis and Palestinians *must* avoid unilateral acts and declarations [...]

and also because they can be 'reproof' words, can be used to rebuke one's opponents and critics:

(3) MR LOCKHART: [...] if that's all he has to offer to this debate, he ought to stay out of it.

0.3.2 Hiding information

The modality system allows speakers/authors to be as precise as they want to be. The transitivity system often forces people to be *more* precise than they might want to be (see Section 0.2.3). But language also offers a number of techniques for being vaguer or for giving *less* information than one might.

0.3.2.1 The absent quantifier

The first of these is the *removal of the quantifier* from a noun phrase, beloved of journalists (and just perpetrated by the present writer). Examples are not hard to find:

Angry voters in the married barrister's Essex constituency urged him to quit [...] And ministers were counting on the local party to order him out.
(*The Sun*: in Reah 1998: 71)

How many voters, how many ministers – all, most, a few, two? The writer either could not or did not want to be more precise, and as long as at least two members of the categories in question want the poor barrister 'out', he or she cannot technically be accused of lying. This can be a highly misleading and pernicious argumentative ploy. In particular, it can lead to prejudicial over-generalization of minority (or even majority) groups, for example, *blacks are violent, men are rapists* and so on, because the hearer/reader can be left to understand the missing quantifier in such cases as 'all', on the analogy of *mountains are high* or *cats are feline*.

To study this phenomenon, one might prepare concordances featuring the significant protagonists in the texts one is studying. To cite an example regarding the briefings corpus, the item *Yugoslavs* appears six times in the briefings corpus, all but one of these in journalists' turns, and the one time the word is uttered by the podium it is in echo of a question:

(4) Q: You don't really expect the Yugoslavs and Milosevic to agree to the Rambouillet accords at this stage of the game, after the bombing, do you?
MR LOCKHART: The Yugoslavs and President Milosevic needs to accept the essence of those accords

In contrast, the podium often speaks of *Americans*, for example, 'things that Americans recognize', 'the long-term benefits Americans have come to expect'. He is, presumably for diplomatic reasons, very careful not to over-generalize and

speak of (all) *Yugoslavs*, but is often quite ready to assume the voice of (all) *Americans*.

0.3.2.2 Nominalization

The second information-impoverishment technique is nominalization, the representation of a process as a noun. This is a classic example of what Halliday (1995: 321) calls a 'grammatical metaphor', that is, the presentation of what in nature would be one kind of phenomenon in language as quite another. It is characteristic of all adult language and indeed is so pervasive that, like many other forms of metaphor, is barely considered by speakers as being metaphorical at all, for example, 'Bad weather has delayed many trains this morning. *Delays* are particularly severe on the east coast'. It functions as an information-packaging or *encapsulation* device (Thompson 1997: 170), especially when referring cohesively to events already mentioned, and as Halliday himself notes is particularly frequent in some genres, such as scientific discourse (Halliday and Martin 1993).

The characteristic of nominalization which interests us here, however, is that it removes the indications of time and modality that are generally present in a verb clause. It can also remove participants in a process:

These ideas have been subject to widespread *criticism*
The *coming* of writing.
(Thompson 1997: 167)

Who has criticized these ideas? Where did writing come to? Sometimes these can be recovered from context or by using one's knowledge of the world. Alternatively, they can be reintroduced in some other part of the utterance. But if they are not they can be lost from view or hidden from view. As Fowler remarks:

Nominalization [...] offers extensive ideological opportunities [...] To understand this, reflect on how much information goes unexpressed in a derived nominal, compared with a full clause: compare, for example, 'allegations' with the fully spelt-out proposition 'X has alleged against Y that Y did A' [...] Deleted in the nominal form are the participants (who did what to whom?), any indication of time – because there is no verb to be tensed – and any indication of modality – the writer's views as to the truth or the desirability of the proposition.
(Fowler 1991: 80)

Fairclough (1989: 124) also discusses nominalization as one of the features of grammar that is particularly powerful in building a picture of the world (which has, in his terms, 'high experiential value').

Nominalization in given texts can be studied, like other metaphorical forms, by examining the frequency lists that derive from them, especially the relative frequency lists (keyword lists, see Section 0.4.1), which tell us which words are more frequent

in the types of text we are examining. A comparison of White House briefings with a corpus of news interviews and another corpus of written news texts revealed a number of interesting aspects of nominalization use. They included, for example, *operations* and *strikes* which, like other modern euphemistic military vocabulary, can be used to render war victimless. We find *air operations, field operations* and *NATO strikes*. Only infrequently do we have the full complement of protagonists, including the goal, appearing in the phrase, as in *NATO air strikes against Yugoslavia*.

Another example is illustrated by items like *cooperation, agreement*, which are not always accompanied by indications of who the parties agreeing/cooperating are or what they are agreeing to do or cooperating in:

(5)   Q: [...] It sounds like there's a lot of convergence here, that military *cooperation* can be enhanced without getting into this definitional rankle.

Note also the nominalized term *convergence*. The press are in fact very aware of the indeterminacy of these items. They frequently ask the podium to define what is meant by *agreement* (the long hyphen indicates an inaudible passage):

(6)   Q: Joe, can I ask one more question on this *agreement* definition? Can *agreement* be implicit so that if, say, Serbian forces are basically destroyed – opposition that there's an implicit *agreement* that U.S. ground troops are allowed?
MR LOCKHART: Again, I don't know how the *agreement* manifests itself, but I'm not sure that there can be an implicit understanding.

Agreement, convergence, understanding and *cooperation* have been reified, have practically reached a status of independence from human beings doing the agreeing, etc. *Agreement, cooperation*, etc. are seen as 'good things' in themselves. We should probably be wary of reading too much that is sinister in this use of language – the nominalization of such items is fairly typical of political talk. Having said this, it does seem that *agreement* in the context of the Kosovo crisis tends to really mean Yugoslavia giving in. This is more explicitly expressed with the nominalization *compliance*, also present in the keyword lists.

0.3.2.3 *Adjectivization*

The third linguistic means of hiding information is *adjectivization*, that is, the employment of nouns as modifiers of other nouns. This is also a type of grammatical metaphor but one that has received much less attention than nominalization. And like the latter, adjectivization is such a standard feature of the language that we are rarely aware of the possibilities for choice and of the ambiguities it can allow. Consider a phrase like *the American President*, which means both 'the President of the Americans' and 'the President who is an American'. The ambiguity is not important in most contexts. But in, say, *Muslim leader*, the

descriptive term can also mean something more than just belonging to a nation or ethnic group, but being a *real, proper, authoritative* Muslim.

We might also reflect on sentences such as:

(i)   AIDS is a gay problem.
(ii) BSE is a British problem.

Do these mean:

(a)   AIDS/BSE is a problem *of* (besetting) gay/British people;
(b)   AIDS/BSE is a problem (caused) *by* gay/British people;

or

(c)   AIDS/BSE is a problem *for* gay/British people *to solve*?

The adjectivization of *gay* and *British* hides the relationship between the group and the *problem* which, instead, is apparent when a preposition is present. Again, this can offer substantial ideological opportunities. A speaker who was both homophobic and hypocritical could use sentence (i) to mean (b) that AIDS is *caused by* gay people – but, if challenged, could instead claim they only meant (a), it *afflicts* them.

In the briefings we find the following:

(7)   Q: Joe, one of the reasons that you didn't want this conflict to spread is because, if there were mass amounts of refugees, it would destabilize the region. And now that there are [...] do you feel that you can handle *the refugee problem* indefinitely, and prevent it from becoming a destabilizing force?

What is meant here by *the refugee problem*: besetting the refugees or caused by them? And, if the latter, a problem for whom? Probably aware of the particular semantics of the term *problem*, the podium prefers to use *issue* and *question*.

Since they are noun-like in appearance, adjectivized items are not immediately obvious in frequency lists. In the briefings frequency lists, however, a number of words appeared as possible candidates, for example, *ground* (collocates with *forces* and *troops*), *security* (*advisor, arrangements, force, guarantee* and several more – but whose *security?*), *defense* (*concerns, spending*), *air* (*campaign, defenses, power, strikes*) and *peace* (*agreement, deal, process, settlement, talks*). All the above appeared in at least 50 per cent of cases in modifier position.

If one is interested in researching these items for their own sake, one or two heuristics can be used to spot likely areas. Nationality or group nouns, especially if found in singular form, for example, *American, Irish, Catholic* generally modify other nouns. It is worth concordancing items like *problem, issue, question, concern, situation* and so on to see how they are modified. A very good place to find them

is, of course, in newspaper headlines (see Morley 1998 for a detailed discussion of extended noun phrases in this genre and especially of how to interpret the relations between their components).

0.3.3   *Semantic prosody*

A third way in which corpus technology can help expose speakers' and writers' attitudes is in uncovering choices of *semantic prosody* they might have made. Some words regularly collocate with items of a favourable or unfavourable connotation and are said to have a 'good' or 'bad' semantic prosody. The item *prosody* is borrowed from Firth (1957), who uses it to refer to phonological colouring that spreads beyond segmental boundaries. Here, meaning is spread beyond word boundaries since it belongs to both the word and its collocate. An example I have used elsewhere (Partington 1998: 67) is *rife*. The denotational meaning of this word is simply 'of common or frequent occurrence, prevalent' (Webster's Encyclopedic Dictionary 1989). But, as the following concordance (*Newspaper corpus*) reveals, it collocates almost exclusively with unfavourable items, such as *crime, diseases, corruption* and *violence*.

```
 1  They are seedy, run-down areas where crime is rife and the misery of unescapable poverty stalks
 2  when rickets and other deficiency diseases were rife. He stayed with Glaxo all his working life,
 3  to be managing director. Uncertainty is already rife. Several months ago, when rumours circulated
 4     in an economy void of privatisation laws and rife with corruption. Delegates also voted to defend
 5     chumminess is more popular, so mistakes are rife. The change has been gradual since the Second
 6     ding societies ripe for takeover. Speculation is rife in the building society industry. Patrick Hosking
 7     management group last month. Speculation is rife that it will soon sell the reconstituted Lehman
 8  n to democracy: Vote-buying and violence are rife as Thais prepare for polling day, writes Teresa
```

The phenomenon becomes especially interesting when people diverge from the expected profiles, when they upset these normal collocational patterns. Louw (1993) argues that this can be done consciously, in search of ironic effect. A phrase like 'conservatism is *rife* in middle England' would almost certainly be so heard.

But Louw also argues that writers can also diverge from a prosody by accident, in which case the reader may detect a difference between what the speaker or writer is apparently saying and what he/she really believes. I found several examples in these briefings, for instance:

(8)   MR LOCKHART: I think General Clark has answered that question, which is he could *deal with success* quite quickly.

The concordance of *deal with* in these briefings shows it very often collocates with unpleasant items: *aggression, situations of conflict, the scourge of terrorism* amongst others. If we take Louw's argument that a speaker's unconscious upsetting of normal prosody may reveal something of their real attitudes, that of the phrase *deal with*

success may be telling us that the podium (or General Clark) is in his heart fearful about the prospects of success. In addition, the kinds of *people* whom the podium and his clients have to *deal with* are usually enemies or, at the very least, problem groups of some type. They include *Milosevic, the Belgrade authorities, China*. We might then be surprised by Mr McCurry's reply to the question of who the podium works for:

(9)   So you work for both sides of this equation. I like to tell people, my office is perfectly situated as a geographic metaphor here in the White House – 50 feet in one direction is the Oval Office, and 50 feet away is here where we are *dealing with you*.

He claims to be working as much for the press as the President but the phrase *dealing with you* suggests that he really sees the press as a problem not a client.

Another example of misfiring prosody is:

(10)   MR LOCKHART: President Milosevic has proved to be *quite adamant in his campaign of atrocities* in Yugoslavia and in his willingness to pay a price.

The concordance of this item from the *BNC* shows that whoever is *adamant* has strongly held beliefs and convictions. In general, it is fairly honourable in its connotations, for example:

1   Lawyers are *adamant* about the need to remain independent of the state.
2   Joanna's grandfather was also *adamant* she was telling the truth.

To be *quite adamant* in a *campaign of atrocities* is odd on a number of levels. The podium has other words at his disposal: *stubborn* or *pig-headed* to express Mr Milosevic's determination; *unscrupulous* or *Machiavellian* to express his supposed villainy. Following Louw's reasoning once again, does the podium's use of *adamant* reveal a suspicion that Mr Milosevic may, by his own lights, have a case?

Without help, it is very difficult to study prosodic effects. Native and non-native speaker analysts may instinctively *suspect* items are being used in unusual ways, but without being able to concordance a corpus to look at the usual patterning, it becomes impossible to prove anything. Both groups can profit equally from corpus technology.

0.4   The importance of comparing corpora

It is impossible to over-emphasize the philosophical point that all work with corpora is properly comparative. Even when a single corpus is employed, it is used to

test the data it contains against another body of data. This may consist of the researcher's intuitions, or the data found in reference works such as dictionaries and grammars, or it may be statements made by previous authors in the field. In the case of language students, the other body of data might be what they have learnt from some authority, such as their textbook or teacher.

Such use of corpora is, therefore, clearly subversive, but it is also in line with empirical methodology. The testing of intuition and/or received knowledge against a body of data is the very spirit of scientific experimentation. Moreover, observations from a single source (even an authentic text) are of limited value and are essentially anecdotal: 'by and large, we are not methodologically justified in interpreting the significance of a particular linguistic event unless we can compare it with other similar events' (Partington 1998: 146). Testing observations and findings against corpus data can provide 'background information' against which particular events can be judged, as we shall see several times in the course of this book.

A word of warning is, however, in order. The corpus should never be treated as the final word. All corpora have limitations, imperfections, *lacunae*. No corpus is fully representative of the language as a whole, or even a subset thereof.[8] In addition, the process of comparison, as we will see, is fraught with dangers. There has been a tendency in some branches of corpus linguistics to raise 'the corpus' to the status of the new authority. But all findings resulting from corpus study should, wherever feasible, be tested anew, very possibly using other corpora.

A good deal of language research, therefore, avails itself of more than one corpus. For translation studies (as well as translation practice), it is clearly useful to have at one's disposal similar or comparable or parallel corpora in the languages of interest (Teubert 1996; Aston 1999). Similarly, it is often profitable to compare sub-sections of the same language, for example, different registers (Biber 1988), the language of different historical periods (diachronic studies) or of different geographical areas (e.g. the English used in United States, United Kingdom,[9] Australia, India and so on) or different text-types, as here.

0.4.1   *The means of comparison: frequency, keyword lists and concordancing*

As implied in the previous section, simple frequency of occurrence of an item in a corpus is not especially informative: the fact that *you*, for example, occurs more often than *therefore* tells us very little. However, should *you* occur significantly more frequently in one corpus than another (say once every 100 words as opposed to every 300), this may well imply something (probably that the first comprises spoken, the other written texts). One might suspect that if, of two written corpora, one contains *therefore* significantly more often than the other, it probably consists of texts of a more formal register, although nothing can be inferred with certainty from a single item.

WordSmith Tools, a suite of programs for language analysis designed by Mike Scott at Liverpool University, provides a *Wordlist* facility that can be used to provide lists of the most frequent lexis in two corpora, one of which is taken as the *foreground* corpus, that is, the one whose language is the object of study, the other the *background* or *reference* corpus. The *Keyword* facility then compares these and produces lists (one alphabetical and one ordered by significance) of all words that are significantly *more* frequent in the first corpus than the second and also of those that are significantly *less* frequent.

Such frequency lists are clearly of their nature blunt instruments but, where appropriate, the clues they provide can be followed up by concordancing the relevant items to get a closer picture of the role they play in briefings texts. In addition, both *MicroConcord* and the *WordSmith Concordance* tool provide a facility that lists the most frequent collocates of a word. This is most useful when examining concordances of very frequent items.

0.4.2   *Choosing corpora for comparison and controlling the variables*

There are, however, a number of practical and methodological pitfalls frequently encountered when attempting to compare corpora.

The main difficulty (and danger) is in controlling the number of variables in play. In choosing a corpus to compare with one's main foreground corpus it is vital, first of all, to compare like with like and, second, to be aware and in control of the differences that inevitably exist between them, in relation to the aim of the comparison. To cite a simple example, if the researcher wishes to uncover differences between varieties differing on the basis of geography, between British and American Englishes, for instance, it would introduce a confusing uncontrolled variable if one of the corpora were of the spoken language and one of the written. If the aim of the exercise is to study differences between the written and spoken forms of the language, a factor of disturbance would be introduced if one of the corpora were of legal texts, the other of doctor–patient interviews, that is, if there were important differences in content. In a *reductio ad absurdum* case, if we compared a corpus of spoken, highly informal, British working class conversations with one of written, highly formal, Canadian government papers, it would be impossible to assign the observed differences in linguistic data to any single cause.

Unfortunately, it is not easy to isolate hermetically the variable to be studied. The general corpora available for commercial or academic use often contain very many different registers but it is not always possible to use them wholesale for comparison. Thus, it is often necessary to compile one's own corpora or tailor those that exist to one's particular purposes, as here.

Even so, the results may still not be perfect. In order to compare the language of the White House briefings corpus (henceforth *WHB*) with another very similar text-type, the news interview, I compiled a separate corpus, and gave it the name

INTS. This contains interviews of political figures conducted on British television by Jonathan Dimbleby (ITV: *The Dimbleby Interview*), John Humphrys (BBC: *On the Record*) and David Frost (BBC: The *Breakfast with Frost* interview) in 1999–2000. It is deliberately the same size as *WHB*, containing around 250,000 words. This genre was chosen in order to attempt to isolate features particular to briefings. For example, since both corpora consist mainly of questions, any differences discovered might tell us something about the particular types of questioning they involve. However, *INTS* is a collection of texts from *British* environments since I was unable to collect sufficient suitable material from US sources. This inevitably introduces a second maverick variable into the comparison. However, attempts were made to control this second variable by checking results against the data in the *Frown* and *Flob* corpora. These are general corpora, a million words each, the first of current American English and the second of current British English, which were designed to be comparable, that is, they contain exactly the same variety of text-types. Thus, if a word or phrase was found to be more frequent in *WHB* than *INTS* (or vice versa), the same word or phrase was looked up in *Frown* and *Flob*, to see if the difference in frequency was the product of a systematic difference between the two language varieties.

A couple of comparative examples should illustrate this point. It was no surprise to find words like *sometime*, *defense*, *gotten* included among the significantly more frequent words in *WHB*, since they are 'American words' or spellings, little used in the United Kingdom (although the last does appear twice in *Flob* in the expression *ill-gotten gains*). However, the keyword lists gave the term *problem* as significantly more common in *INTS* (171 occurrences to 58). But is *problem* simply more frequently used in British English? According to the comparison between *Frown* and *Flob*, the answer is no: *problem* was actually *more* common in the American texts (296 to 217). On the other hand, *solution* was four times more common in *WHB* than in *INTS* (44 to 11), whilst their occurrences in *Frown* and *Flob* were roughly equal (65 to 69). Thus, the use of these items is probably saying something about these sets of texts, not about American or British English. *Problems* greatly outweigh *solutions* in the news interviews, whereas there is just about one *solution* for every *problem* in the briefings. The podium, in fact, prefers to use the words *issue* or *concern* – both of which are in the keyword list. The podium is clearly keen to portray an optimistic vision of events, whereas many of the news interviewees, perhaps hostile to the government of the day, are happy to talk at length of any *problems* it may be encountering.

The picture is more complex when whole sets of items are found to be more frequent in one of the corpora. *INTS* is significantly richer than *WHB* in a large number of adverbs, many of which are intensifiers: *perfectly*,[10] *extremely*, *absolutely*, *indeed*, *very*, *really* and *quite*.[11] From this cue, a second look at the keyword data, along with subsequent concordancing of likely items, also showed that a fair number of other items frequently used to intensify were much more common in the news interviews: *whole* (found e.g. in 'a whole new system'), *huge* ('huge work', 'a huge task'), *all* ('all across the country'), *everybody* ('everybody knows

that ... '), *nobody* ('there is nobody more determined ... '), *always* ('I always argue that ... '), *never* ('he never said any such thing ... '), along with *much*, *more*, *big* and *better*. Of all these, only *perfectly*, *indeed* and *quite* were also found to be more frequent in *Flob* than *Frown*. A high degree of intensification, of hyperbole, would seem to be a feature of these interviews in relation to the briefings. This is somewhat surprising given, first of all, the folk reputation of American speakers as being more expansive than the British and, second, the temptation for the podium to exaggerate his clients' virtues. However, as we shall see (Section 7.5.2), there are brakes on his boastfulness, and overstatement is likely to be mocked by his audience (Section 11.5). There are no such immediate restraints on news interviewees. Moreover, it may also be the case that intensification is a common feature of the kind of language used in *formulating an argument*, in making a case *for* or *against* something or someone, which is by far the principal activity of the political interviewee. A popular tool of political argument is *comparison* and *contrast* and the speakers often wish to stress that their side's approach is *extraordinarily* good and that of their opponents *extremely* poor, for example:

(11)   JIM CALLAGHAN: [...] the government's done *very well indeed*, I think it's been *extraordinarily* good, had good circumstances and done well.

(12)   ANN WIDDECOMBE: Well of course what Tony Blair has done is to be *extremely* cynical about this.

The podium has many other duties to perform besides constructing such arguments – issuing statements, discussing presidential and other business, engaging in banter with the members of the press and so on, and, therefore, features of argumentative prose, including intensification, are relatively less common.

There can often, then, be considerable difficulties in finding the 'perfect' comparable corpus and in controlling the number and degree of differences. Moreover, for our purposes, it would scarcely be advisable to measure one corpus against another that was too *similar*. For example, comparing *WHB* to another corpus of briefings, perhaps from a different source or a different period of time would probably reveal very little about the essential nature of briefings, about their linguistic peculiarities. The keyword lists would only throw up lexis highlighting different topics of discussion and different personalities of the day, everything else, being the same, would be invisible.

0.4.3   *Other pitfalls in comparing corpora*

The word and frequency lists used in this study are unlemmatized, that is, the machine does not distinguish between different words written the same way, or between different forms of the 'same' word (say noun and verb) if they are written the same way.