Corpus-Assisted Discourse Studies on the Iraq Conflict

Wording the War

Edited by John Morley
and Paul Bayley

ROUTLEDGE

Corpus-Assisted Discourse Studies on the Iraq Conflict

Routledge Advances in Corpus Linguistics

Edited by Tony McEnery, *Lancaster University UK*
Michael Hoey, *Liverpool University, UK*

1. Swearing in English
Bad Language, Purity and Power from 1586 to the Present
Tony McEnery

2. Antonymy
A Corpus-Based Perspective
Steven Jones

3. Modelling Variation in Spoken and Written English
David Y. W. Lee

4. The Linguistics of Political Argument
The Spin-Doctor and the Wolf-Pack at the White House
Alan Partington

5. Corpus Stylistics
Speech, Writing and Thought Presentation in a Corpus of English Writing
Elena Semino and Mick Short

6. Discourse Markers Across Languages
A Contrastive Study of Second-Level Discourse Markers in Native and Non-Native Text with Implications for General and Pedagogic Lexicography
Dirk Siepmann

7. Grammaticalization and English Complex Prepositions
A Corpus-Based Study
Sebastian Hoffman

8. Public Discourses of Gay Men
Paul Baker

9. Semantic Prosody
A Critical Evaluation
Dominic Stewart

10. Corpus-Assisted Discourse Studies on the Iraq Conflict
Wording the War
Edited by John Morley and Paul Bayley

Corpus-Assisted Discourse Studies on the Iraq Conflict

Wording the War

Edited by John Morley and Paul Bayley

Routledge
Taylor & Francis Group
New York London

First published 2009
by Routledge
270 Madison Ave, New York, NY 10016

Simultaneously published in the UK
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

This edition published in the Taylor & Francis e-Library, 2009.

To purchase your own copy of this or any of Taylor & Francis or Routledge's
collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk.

Library of Congress Cataloging in Publication Data

Corpus-assisted discourse studies on the Iraq Conflict : wording the war / edited by John Morley and Paul Bayley.
p. cm.—(Routledge advances in corpus linguistics ; 10)
Includes bibliographical references and index.
1. Discourse analysis—Political aspects. 2. Corpora (Linguistics) 3. Iraq War, 2003—Language. I. Morley, John, 1940 Dec. 17–
P302.77.C68 2009
401'.41—dc22
2009015185

ISBN 0-203-86815-3 Master e-book ISBN

ISBN10: 0-415-87137-9 (hbk)
ISBN10: 0-203-86815-3 (ebk)

ISBN13: 978-0-415-87137-2 (hbk)
ISBN13: 978-0-203-86815-7 (ebk)

Contents

Concordances

Figures

Tables

Acknowledgments

The editors of this volume would like to thank the Italian Ministry of Education, University and Research for the grant which allowed this project to be realized. They would also like to give their warmest thanks to their colleague, Guy Aston, whose expertise as well as his unceasing and meticulous work created the corpus on which the project was based, and whose constructive criticism kept all the contributors on their toes.

The editors also thank: J. R. Martin, P. R. R. White, and Palgrave Macmillan for permission to print Figures 2.1, 2.8, and 2.9, originally published in *The Language of Evaluation: Appraisal in English*, 2005; Guardian News and Media for permission to print in Chapter 10 the article "Civilian Targets", which appeared in the *Guardian* 14/04/2003, copyright Guardian News and Media Ltd. 2003; and finally the *Sun*, for permission to print in Chapter 10 the article "Good Evening: Here Is the Worst Possible News", published on 11/4/2003 under the byline Littlejohn.

Introduction
A Description of *CorDis*

John Morley

0.1. GENESIS OF THE BOOK

This book is the product of the collaboration of a group of scholars in different Italian university institutions which started in 2004.[1] We had a common interest in corpus linguistics, discourse analysis, and institutional and media discourse, though we had all been pursuing rather different lines of research. In 2003 the group put together a bid to the Italian university funding body for a research project which combined our interests. It had the official title *Corpora and Discourse: A Quantitative and Qualitative Linguistic Analysis of Political and Media Discourse on the Conflict in Iraq in 2003*, and became known by the acronym *CorDis*.[2]

Two main elements helped to unify our research: the first was that we were all working on a corpus of texts concerning the Iraq war, the *CorDis Corpus*; and the second was that we were all committed to a methodology which was known within the group as corpus-assisted discourse studies (CADS). We will say a little more about this methodology later. The chapters in this book all derive from research on this corpus and to a greater or lesser extent use the CADS methodology. CADS, like all studies involving corpus linguistics, builds on the work of Sinclair, Hoey, and Stubbs and their names will be mentioned frequently in the volume (in particular, Hoey 2005; Sinclair 1991, 2004; Stubbs 1996, 2001). We would like to pay a tribute to our recently deceased colleague, John Sinclair, to whom all members of our research group owe an enormous professional debt and to whom some of us were linked by ties of friendship. Another unifying factor was that most of the group had a commitment to systemic functional grammar (see, in particular, Halliday 1994), which is the grammatical framework that predominates in the analyses in this book; there was also an interest among members of the group in exploring aspects of stance and evaluation, that is, how speakers and writers 'instruct' their interlocutors on how to interpret their messages (see Hunston and Thompson 2000; Martin and White 2005).

0.2. THE COMPOSITION OF THE CORPUS

The *CorDis Corpus* consists of six related projects. The projects are briefly presented here in an order which reflects an ideal temporal progression. The acronyms refer to the subcorpora which the projects produced. Note that the newspaper project produced three subcorpora corresponding to news stories, editorials, and op-eds.

- Sources of news creation—(a) British House of Commons (*HoC*), (b) US House of Representatives (*HoR*) (Projects 1 and 2).
- News negotiating and mediation—White House press briefings (*WHB*) (Project 3).
- Recounting to the public—(a) television news (*TVNews*), (b) newspapers (*PapNews*), (*PapEds*), (*PapOp*) (Projects 4 and 5).
- Parliamentary inquiry—Hutton Inquiry (*HUTTON*), which touched on the British Government's reasons for going to war (Project 6).

The way the corpus has been marked up allows us to create different groupings or *classes* of texts in order to answer specific research questions. These classes may cut across subcorpora, for instance, all the texts from US papers, whether they are reports, editorials, or op-eds; or they may be parts of one subcorpus, such as the information-gathering section of the Hutton Inquiry subcorpus. We refer to the different ways of dividing up the corpus into classes as *partitions*.

0.2.1. Discourse of the Legislative Assemblies (Projects 1 and 2)

These two projects derive most immediately from work demonstrating that much political action is constituted by linguistic action and as such is a legitimate field of study for the linguist (see Blommaert and Bulcaen 1998; Chilton, Ilyin, and Mey 1998; Fairclough 1995; Geis 1987; Wilson 1990; Wodak 1989). Parliamentary discourse can be considered as prototypical political language and yet it differs from much of what we now call political language because it can only take place in one institutional arena, and in order to participate in it one has to be an elected member of the institution. Previous volume-length studies of parliamentary discourse, which provide a starting point for the project, include Bayley (2004); Carbò (1996); Wodak and van Dijk (2000). This research has shown that a cross-cultural analysis of parliamentary language can be extremely fruitful because on the one hand parliaments in Western democracies fulfil, in and through language, similar functions—they legitimate and/or contest legislative proposals and policy orientations, they subject the executive power to scrutiny, and they represent constituency or other interests—but on the other they differ in terms of their rules and regulations, their representativity, and their accountability. In particular, they are expressions of different

political cultures—long-term orientations towards government and general beliefs, symbols, and values (Heywood 2000). The two chapters (2 and 3) by Miller and Johnson and Bayley and Bevitori seek to identify how this political culture in different but allied nations (the US and the UK) is articulated in the discourse of parliamentarians justifying or contesting military intervention.

0.2.2. The White House Press Briefings (Project 3)

The White House press briefings project deals with an extremely recent linguistic-political-media discourse type which evolved in the 1990s in the United States from press conferences (Clayman 1993). Being a new genre, very little has been written about these briefings; indeed, previous work has focused on the genre as a site of political action: Maltese (1992) and Kurtz (1998) have examined, for instance, how the US government attempts to 'spin' its message in times of conflict and the press's reaction to such attempts. Partington's (2003) book-length study is the first work to approach this discourse type using a full range of linguistic tools. One of his main points is that the briefings represent an excellent site for the study of the evolution of a new discourse type. Furthermore, he addresses in some detail how the participants invent ex novo the rules of a novel interactive 'role-play'; and how they learn to behave both in cooperation but also in competition (the journalists with the president's spokesperson, the podium—and vicariously with the podium's political masters—and vice versa) with the other participants. Riccio's chapter is an example of how a seemingly neutral word, *message*, is spun so that it takes on connotations of menace.

The briefings are, in fact, frequently the arena where White House policy is first aired—sometimes even before it has actually officially been formulated. Moreover, although they are ostensibly a kind of mediation, whereby the White House states its agenda and the press decides how it is going to report it, briefings are also in effect a way for the White House to get its message over the heads of the press directly to ordinary citizens.

It is interesting to note that many major US news outlets have a regular section reporting what goes on in briefings, and this makes the podium a highly recognisable media-political celebrity in his own right.

0.2.3. TV News Discourse (Project 4)

The TV news discourse project focuses on the language of television news and starts from the recognition that, like the rest of the media, TV news is involved in creating what Hall et al. (1981) call 'maps of meaning': that is, the presentation of news to the public in ways that they will understand. As Galtung and Ruge (1981) have it, what is signal and what is noise is not inherent; it is a question of convention. MacDougall puts this more philosophically,

> [J]ournalists do not gather news; they construct second order accounts of reality from materials provided by sources (first order accounts).
>
> (MacDougall 1983:85–6)

The creation of these conventions, which we all recognize, means that the news stories must conform to criteria of newsworthiness that are accepted both by the news creators and their audience. It could be argued further that TV news 'operates within the framework of the dominant value system and therefore helps to maintain the *status quo*' (Selby and Cowdery 1995:144). This aspect of news reporting is central to the television news project, which investigated the media construction of the conflict in a comparative perspective across four networks representing public and private channels in the UK, US, and Italy.

Television news also, perhaps predominantly, makes use of images. Although images are not the direct object of study of either of the chapters concerning television in this volume, as they are in Lipson's essay in our sister volume, they are the background against which the linguistic analysis is performed. [3] Here we acknowledge a debt to the pioneering semiotic media work of Fiske (1987) and Hartley (1982).

It is perhaps strange that so little work has been done on the linguistic realization of the second-order accounts of the world presented in television news. A few exceptions might be represented by Iedema, Feez, and White (1994), Fairclough (1995b), Haarman (1999, 2006), and Lombardo (2001, 2004). The chapters by Lombardo and Clark (Chapters 5 and 6), which explore the television data using slightly different analytical approaches, have the virtue of representing a systematic study of this genre using an extended corpus of texts: they both deal with US and UK television news programmes.

0.2.4. Newspaper Discourse (Project 5)

The newspaper discourse project, too, starts from a similar acknowledgment of the complex relationship between events and their presentation through the media. As Chibnall says,

> The reporter does not go out gathering news, picking up stories as if they were fallen apples, he creates news stories by selecting fragments of information from the mass of raw data he receives and organising them into a conventional journalistic form.
>
> (Chibnall 1981:76)

The work of Fowler has exercised considerable influence on this part of the research. In his study of discourse and ideology in the press, he states 'my major concern is with the role of linguistic structure in the construction of ideas in the press' (1991:1). Another fundamental text for research on

newspaper discourse is Bell (1991) on the language of news media, which combines the insights of a linguist with the experience of a working journalist, as does White (1997), whose work on distinguishing journalistic discourse types is an important basis for further research. Morley (2004a, 2004b) and Murphy and Morley (2006) have looked at the difference between news stories, editorials, and op-eds. This is the aspect of the research which is followed up in Murphy's chapter on newspapers (Chapter 7).

0.2.5. The Hutton Inquiry (Project 6)

Like the projects on parliamentary discourse and presidential press briefings, the Hutton Inquiry project deals with spoken language in an institutional context, an area of linguistic study set out in systematic form in Drew and Heritage (1992). In July 2003 the British prime minister appointed Lord Hutton to head an inquiry into the circumstances surrounding the death of Dr David Kelly, a scientist working for the government in weapons inspection. A BBC journalist, Andrew Gilligan, claimed that an unnamed source, who was later discovered to be Dr Kelly, had told him that information contained in an intelligence report had been "sexed up" in a government document justifying going to war. Dr Kelly was caught in the cross fire between the BBC and the government and committed suicide under the strain. (cf. note 5 of Chapter 9 for more details.). Here we were working with official transcripts made available on the Hutton Inquiry Web site, http://www.the-hutton-inquiry.org.uk/. The inquiry was directed by Lord Hutton and offers the researcher the possibility of comparing two related but fundamentally different discourse types—information collecting and adversarial probing of the information supplied. Texts which form the basis of research on these discourse types are Grimshaw (1990), Hutchby (1996), and Partington (2003), all of which examine the discourse involved in conflictual situations, particularly in political and institutional contexts. A linguistic analysis of these inquiries presents the opportunity for Taylor to examine a hitherto rarely explored discourse type (Chapter 8).

0.3. MARKUP

The *CorDis Corpus* is a multigeneric corpus containing over five million tokens and about fifty thousand types of both writing and transcribed speech. The corpus brings together eight subcorpora or modules. The construction of a modular corpus brings some practical advantages in that each module can be used independently and could at some moment in the future be added to at will (see Haarman et al. 2002). The corpus was marked up by a team based at the University of Bologna's School for Interpreters and Translators in Forlì, overseen by Guy Aston, as a set of Extensible Markup

Language (XML) documents which conformed to the Text Encoding Initiative (TEI) guidelines. It was designed to be interrogated by *Xaira* (XML Aware Indexing and Retrieval Application), a software developed by Burnard and Dodd at Oxford University Computing Services (see the Web site http://www.xaira.org). The *CorDis Corpus* served as a one of the test beds for the development of that software. It was tagged for part of speech and lemmatized by Rayson's group at Lancaster University, using the CLAWS7 tagset (see Rayson and Garside 1998).

Initially, the markup was seen as being simply ancillary to the work of analysis, which was to come after its completion. In practice, we found that marking up the text was in itself a form of text analysis. This is particularly true of the sections of the corpus containing the *BBC* and *CBS* news programmes. There were no ready-made categories existing for the structure of TV news in the TEI guidelines, and so our divisions had to be based on work already done by members of the research team, in particular by Haarman (see the chapters by Lombardo and Clark on TV news in this volume).

0.3.1. TV News Markup

Some of Haarman's markup codes can be applied to any television news programme;[3] others are specific to the news programmes in the TV news subcorpus which is part of the *CorDis Corpus*. First of all, each news report is indicated as a separate section by the markup. Headlines are then identified and coded; if they appear on the screen, they are also coded for that. Different speakers' utterances are identified, with markup to indicate the identity, sex, and role of the speaker. The default situation is that speakers talk to the camera. If s/he speaks in voice-over, this too is indicated. When the newsreader introduces a reporter's report, this is also indicated in the markup. The reporter's report also includes coding for his/her precise role, whether s/he is a:

- studio reporter
- embedded reporter
- war zone correspondent
- correspondent (e.g., from Baghdad, Washington, Brussels)
- reporter plain and simple.

Different parts of the utterance are marked up as one of the following:

- text spoken by the reporter over video actualities
- text spoken by the reporter to camera
- text spoken by the reporter via telephone link.

Apart from the reporters and studio presenters, other speakers too are identified by their functions, either as:

- legitimated persons: that is, speakers who have the status to speak for others because of their status, e.g., politicians, professors, doctors, experts of some kind
- *vox populi*: that is, members of the public, normally unnamed
- military: that is, members of the military who have not sufficient status to count as legitimated persons.

Where relevant, addressees of utterances are indicated (such as questions asked by a reporter to a legitimated person or a *vox populi*, or by a newsreader to a reporter, or vice versa).

All the codings involved decisions about the importance of these sections of text for the analysis of a television news programme and are the result of research questions formulated by the colleagues working on television news programmes.

Although a considerable amount of time and effort, and a large amount of our research funding, was dedicated to marking up the corpus so that it could be interrogated by the *Xaira* software, we did not abandon *WordSmith Tools*, the software that most of us had 'grown up' with. There were a number of reasons for this, apart from the comfort of familiarity: first of all, we were anxious to get on with our analyses as soon as the corpus existed in an exploitable form and did not want to wait for the lengthy process of the XML, TEI conformant markup to be completed—about forty presentations and papers have been produced in the three years since the project began, many of which will be cited by the authors of the various chapters. Secondly, *WordSmith Tools*, both version 3 (Scott 1999) and version 4 (Scott 2005), produce instantaneous word lists and keyword lists, which are often the birthing point of research questions. And finally, as we were all working with relatively modestly sized subcorpora, on average about a million tokens, of text-only files, it was sometimes possible to add what our markup experts call 'light markup' (see Chapter 1), tailor-made for individual research questions.

A number of corpus linguists, notably Sinclair (2004:190–1), have argued that some forms of markup condition and prejudice later researchers' exploration of the corpus, but we hope to 'use tags en route to the language, and not just stop there', to quote his own words (2004: 191). We believe, as Cirillo, Venuti, and Marchi argue in Chapter 1, that markup favours replicability and enhances the reliability of the research. Careful markup of the rhetorical structure of texts certainly aids the work of comparison between these different parts of the discourse structure.

0.4. COMPARISON

Corpus-assisted discourse analysis, the kind of corpus linguistics embodied in this project, of necessity entails comparison, both at a fundamental

ideological level and also in many methodological-practical ways. In general terms, the statement that any given linguistic feature being studied is frequent or infrequent in the discourse type contained in corpus X only has proper contextual significance when corpus X is compared to corpus Y, which normally contains another discourse type. The choice of corpus Y, the comparison or background corpus, needs to be carefully made. We may want to compare a specialized discourse type against general English, in which case our Y corpus will be one of the large corpora of general English, such as the *BNC*. We may, on the other hand, wish to compare one kind of specialized discourse with another kind of, perhaps superficially similar, specialized discourse: we may, for instance, be asking if editorial articles (corpus X) differ from op-ed articles (corpus Y) (see Murphy, Chapter 7). Or we may be interested in diachronic change: it could be that we want to know if White House press conference language has changed since before 9/11 (Riccio, Chapter 4). In this case the X and Y corpora will be of the same discourse type but from different historical moments.

One thing which markup clearly does is allow the researcher to identify quickly and efficiently subparts of a corpus and compare them against one another. We might illustrate this by looking at the newspaper subcorpora. It was decided to divide the newspaper section of the *CorDis Corpus* into three subcorpora—news reports, editorials, and op-eds—because two members of the project group, Murphy and Morley, were already conducting research into the linguistic differences between these discourse types. (This is also an illustration of the observation made by Cirillo, Venuti, and Marchi in Chapter 1 that some elements of the structure of the *CorDis Corpus* were predetermined.)

We then had to decide what other divisions of the newspapers were important for us as researchers. It was fairly obvious that we needed to be able to distinguish the individual newspapers, for instance, the *Guardian* from the *Daily Mirror*. An early piece of research showed differences between the attitudes towards certain aspects of the war of these two newspapers, a left-wing quality and a left-wing popular newspaper (Morley 2005). It was also clear that we wanted to be able to distinguish between UK and US newspapers and between the quality and popular papers as groups, in order to be able to make comparisons between these. As a result of marking up these distinctions it is possible to use the *Xaira* software to make even more precise ad hoc partitions, such as one containing the word *soldiers* in the editorials of US popular newspapers published in a particular week. It would also, for instance, be possible to compare the discourse of all female Democrats with that of all male Republicans. None of us has so far interrogated the corpus in these terms, at least more than informally, but it would be possible and *Xaira* would provide us with elegant histograms or pie charts to illustrate our data.

Another example of the flexibility which *Xaira* affords can be seen in Duguid's work (Chapter 9), where she compares the words which speakers

and writers use to report other discourses across all the subcorpora (the Hutton Inquiry, the White House press conferences, Hansard, the House of Representatives, TV news, and the three newspaper subcorpora—editorial, op-eds, and news reports).

As well as inter- and intra-subcorpus comparison, many of the authors of the book make use of large background corpora. The *BNC*, the new version of which can now be searched by *Xaira* software, is the most commonly used as it represents a large, easily accessible and very reliable corpus of texts which allows us to compare our specialized subcorpora with relatively modern general British English.[4] Partington (Chapter 10), instead, references *SiBol 05*, a collection of more than 150 million words of quality English newspaper texts published in 2005. This was more appropriate than the *BNC* for his research as some of the lexical items he looks at are of relatively recent press coinage.

0.5. CORPUS-ASSISTED DISCOURSE ANALYSIS

Tognini-Bonelli has made an important distinction between 'corpus-based' and 'corpus-driven' linguistics (Tognini-Bonelli 2001:10–11). The corpus-based approach uses the corpus as a library of texts to be searched to test preformed hypotheses. In corpus-driven studies, on the other hand, 'the theoretical statement can only be formulated in the presence of corpus evidence and is fully accountable to it' (Tognini-Bonelli, 11). We have to trust the texts. The first approach illustrates what Ellis (1985) calls 'the theory-then-research approach', or deductive reasoning, and the second 'the research-then-theory approach', or inductive reasoning. This is a very important distinction and it is fairly easy to assign most corpus studies to one approach or the other.

Those of us who adopt a CADS approach would argue, however, that one approach does not necessarily exclude the other. What frequently happens is that we generate a word list, read through it, and our intuition tells us that certain words or clusters are going to be interesting. To give a simple example, I composed a word list of four-word clusters from a half-million-word corpus of newspaper news articles on political reporting. This was done with no idea of what, if anything, of interest would come out. The same was done for a half-million-word corpus of editorial articles from the same period and the same newspapers. The 'key-most' cluster for the reports compared to the editorials turned out to be *for the first time*. My intuition, or rather my intuition primed by years of reading newspapers and about newspapers, immediately suggested to me that this was an interesting cluster: it was an illustration of the scoop mentality of Anglo-American newspapers. The next step was to look at the sixteen instances of *for the first time* and check what their function was in the wider context of the whole article. To recapitulate, then, we have three stages in this research:

(1) the software throws up the clusters *for the first* time as being significantly more frequent in news reports than in editorials, (2) intuition tells us there is reason for this, (3) we check by a close examination of texts to see if our intuition is correct.

A less obvious finding came from the four-word clusters of the editorial corpus: the fifth most frequent cluster was *at the heart of*, with a frequency of fifty-eight per million words.[5] The corpus contained editorials from popular and quality English newspapers printed over a period of two years in twenty-seven different articles, so there was no chance that this statistic was generated by some leader-writer's idiolect; its frequency in the whole of the *BNC* was 9.6 per million words. The cluster is clearly characteristic of editorial writing in English newspapers in the early years of the millennium. To this day I have no idea why. It was a purely serendipitous find (see Partington and Morley 2004 for details of this research).

In general, we can say that CADS methodology is predicated on the belief that the combined use of qualitative and quantitative linguistic analysis is not only possible but that their combined application increases the researcher's analytical capacity to an extent greater than would be predicted from the sum of the two methods. As in all forms of corpus linguistics the concordance line remains fundamental and collocations, which we find normally from scanning the concordance lines, are as Hoey so tellingly puts it, 'both pervasive and subversive' (Hoey 2005:3).[6] However, the bare concordance line strips away most of the context of the original utterance, without which the study of features of discourse becomes problematic. As Biber et al. (1999) say,

> [A]lthough nearly all discourse studies are based on analysis of actual texts, they are not typically corpus-based investigations: most studies do not use quantitative methods to describe the extent to which different discourse structures are used.
>
> (Biber et al. 1999:106)

Our solution is to move backwards and forwards—to shunt, to use a Hallidayan term (1961, in 2002:45)—from the concordance line to the wider cotext. Reading vertically allows one to see patterns, but we also need to read horizontally to arrive more securely at meanings.

The first mention of CADS methodology as such is Partington (2004a), and the final chapter of the current work presents some of his further reflections upon its scientific significance. It builds on the pioneering article of Hardt-Mautner (1995) and has been greatly influenced by the concrete examples put forward by Stubbs in two of his volumes (1996, 2001). We also feel an affinity with the work described by Baker in *Using Corpora in Discourse Analysis* (Baker 2006). The methodology has informed the research of most members of the group for some time now. We believe, however, that the current volume is one of the first works that uses CADS methodology to treat a single theme, in this case the Iraq war, in a book-length study.

NOTES

1. These were the University of Bologna, 'LUISS, Guido Carli' in Rome, and the University of Siena.
2. Ministry protocol number 2004105247.
3. There is a sister volume to this book based exclusively on the TV subcorpora of the *CorDis Corpus*, Haarman, L. and Lombardo, L. (2009) *Evaluation and Stance in War News: A Linguistic Analysis of American, British and Italian Television News Reporting of the 2003 Iraqi War*, London: Continuum. This book also deals with Italian TV data.
4. We had no access to the *American National Corpus*.
5. Biber refers to clusters which occur at least 10 times per million words (0.0001 times per hundred words) as 'lexical bundles' and argues that they often characterize discourse types (see Biber et al. 1999).
6. Partington (2003), speaking of the importance of comparison in corpus work, also calls this methodology 'subversive' when he argues that our students can use the corpus to test 'what they have learned from some authority, such as their textbook or teacher? (p. 20)

1 The Making of the *CorDis Corpus*
Compilation and Markup

Letizia Cirillo, Anna Marchi, and Marco Venuti

This chapter sets out to describe how different subcorpora were integrated to form the unified body of texts known as the *CorDis Corpus*. In particular, it focuses on the process whereby *CorDis* was made an XML-valid, TEI-conformant corpus that can be easily interrogated using *Xaira*. In discussing specific examples illustrating the practice of markup, the chapter highlights the import of annotation as a way to enhance reliability of research and (re-)usability of data.[1]

1.1. INTRODUCTION: AN AIR SCOUT VIEW

The *CorDis Corpus* is a large multimode, multigenre collection of political and media discourse on the 2003 Iraqi conflict.[2] It was generated from different subcorpora previously assembled by various research groups for diverse discourse analytical purposes. A more detailed description of its composition can be found in the introduction.

A significant portion of our work was devoted to making the subcorpora into a unified homogeneously encoded corpus which could be interrogated using *Xaira*.[3] Initially the corpus was only lightly encoded by each research group on the basis of specific research objectives and hypotheses. The heterogeneity of data, the specificity of the genres, and the various methods adopted involved the use of a wide range of coding strategies to make textual and metatextual information retrievable by means of available concordance software. It was clear from the outset that marking up the corpus as a whole would entail various levels of pre-encoded and pre-existing interpretation. The main purpose of this chapter is to show the process of standardization and integration whereby a loose collection of texts has become a stable architecture. The TEI Guidelines proved a valid instrument providing for a hierarchical organization of metadata which makes markup part and parcel of the corpus. We will underline that it is precisely the markup which gives the corpus a sound structure favouring the replicability and enhancing reliability of research.

In discussing examples, we will deal with issues like conformity and validity, and we will examine the constraints imposed on data handling by the methodological framework adopted. In particular, we will argue that the crucial role of annotation leads to a reconsideration of the definition of corpus itself, in which special emphasis is placed on markup being the backbone of the corpus rather than a superimposed accessory.

There is a tendency to distinguish between 'markup' and 'annotation' (McEnery, Xiao, and Tono 2006:29), adopting the first term to refer to contextual information (i.e., editorial and descriptive metadata) and the second to refer to 'interpretative linguistic information'. We will here use the two terms interchangeably, since both notions share the same salient qualities for the purposes of our description: they are both *added value* and they both carry *interpretative information*.[4]

Finally, the fact that markup involves a substantial amount of human intervention on machine-processed data has some crucial implications for corpus-assisted discourse studies (CADS), since it permits the combination of qualitative and quantitative research approaches.

1.2. TERRITORY

We will start by introducing the *territory* of our work, providing a short description of the various components of the *CorDis Corpus* in order to highlight some of the difficulties we had to deal with. This overview of the subcorpora will lead to some mainly theoretical considerations of the role of annotation in corpus design, and to an evaluation of the annotation scheme adopted.

1.2.1. *CorDis*: One Corpus/Many Corpora

As implied in the introduction, *CorDis* is an XML, TEI-conformant, POS-tagged, multimode, multigenre corpus containing over five million word tokens (corresponding to about 50,000 types). It is made up of eight subcorpora of texts from the following sources: British House of Commons (*HoC*), US House of Representatives (*HoR*), White House press briefings (*WHB*), television news (*TVNews*), newspaper reports (*PapNews*), newspaper editorials (*PapEds*), newspaper op-eds (*PapOp*), and Hutton Inquiry (*HUTTON*). Further details about the harmonization of metadata will be presented in sections 1.3.2 and 1.3.3. Here we introduce some of the specificities of these subcorpora in order to illustrate the kinds of issues we have encountered.

The subcorpora include a variety of modes of language use occupying different positions on a written-spoken continuum: official transcripts of speech (Hutton Inquiry, Congressional Record, Hansard, White House press briefings, all of which are heavily edited and adapted for publication in written form), unofficial transcripts of speech (TV news programmes, aiming to

provide an accurate record of what was said without necessarily conforming to strict conventions of published writing), and published writing (newspaper articles); they also include two main geographical linguistic and cultural varieties—American (Congressional Record, White House press briefings, US newspaper articles, and *CBS* TV news) and British (Hutton Inquiry, Hansard, British newspaper articles, and *BBC* TV news). According to the nature of the texts and to particular research objectives, further specification of these text types had to be made explicit. Thus it was necessary to distinguish between different stages of the Hutton Inquiry (Lord Hutton's opening statement presenting the purpose and the structure of the inquiry, taking of witness statements, cross-examinations) and different types of parliamentary proceedings (e.g., Question Time, statements, speeches, and debates—government or opposition-initiated—just to name a few). To take into account all these aspects, we elaborated a series of categorization schemes. Each of these categorizations provides a different way of dividing up the corpus, known as a *partition*: each partition offers a set of classes in which each individual text can be placed.

Each of the subcorpora also posed specific needs related to the institutional or professional context of the texts, their official status, the discourse setting, and also the particular discourse analytic approach taken by the researchers. Obviously all relevant metadata needed to be encoded in order to make them retrievable by means of dedicated software. Using *Xaira*, which will be described in more detail in 1.3.1, it was possible to instantly select as a subcorpus those texts contained in a particular class of a particular partition. Partitions include *mode* (official transcripts, spoken, and written), *origin* (British vs. American), and *source* (see earlier for the full list). A further partition, *specific*, was used to select each specific source (e.g., a single newspaper or TV news programme) or discourse type (e.g., a debate in the Congressional Record or the Question Time in Hansard), as exemplified in Figure 1.1.

Having outlined the main characteristics of the *CorDis* corpus, we can now move to examine some theoretical and practical issues related to the process of annotation.

1.2.2. The Rationale of Annotation: Marking a Path through the Data

Corpus annotation is the 'practice of adding interpretative linguistic information to a corpus' (Leech 1997:2). This definition stresses two fundamental concepts: when we mark up we *add* information, and in modelling this information we are doing *interpretative* work. Annotation is inserted into the text in order to convey meaning and the operation of reflecting, representing, or creating meaning always implies a selection among a series of possibilities. 'Markup *licenses certain inferences* about the text' (Sperberg-McQueen, Huitfeldt, and Renear 2000: online; original emphasis); each selection privileges some meanings over others and therefore marking up is marking a path through the data.

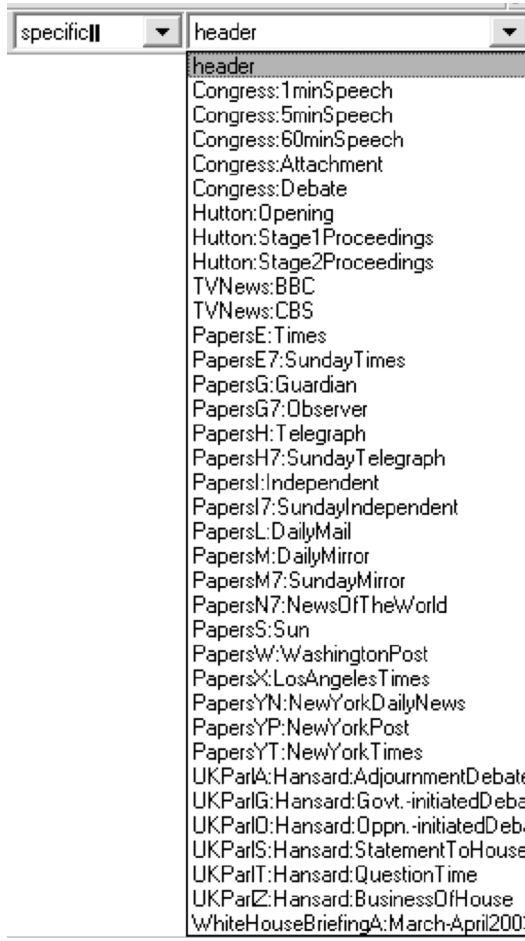

Figure 1.1 *Xaira*'s 'specific' partition for *CorDis*.

Here we argue that annotation, with specific reference to the XML-valid, TEI-conformant markup used in the *CorDis Corpus*, is not merely an accessory to the corpus, a tool that the researcher can use to investigate portions and properties of the texts, but is an intrinsic part of the corpus itself, since it is the annotation that makes the *CorDis Corpus* usable as a whole harmonized and coherent body of texts.

Marking up *CorDis* was crucial for various reasons. These are related both to the general motivations that make the time-consuming task of annotation worthwhile and to the specific nature of the corpus. On the one hand, annotation makes it possible to sharpen the analysis, forcing 'the analyst to test and refine the system of categorization to account for all cases' (Wynne 2005:2), produces more detailed results, and provides a base for replicability of the study and reusability of the data. On the other

hand, the *CorDis* project is based on a collection of heterogeneous texts and text types that finds a common core in its topic (Iraq war in 2003) and in the methods of analysis adopted (CADS) which needs to be harmonized in order to work as an organic apparatus.

'Corpora are useful only if we can extract knowledge or information from them' (Leech 1997:4). Markup is *added value* because it makes built-in information retrievable, allowing the user to access knowledge about the data in the corpus that would be lost if it had not been made explicit in the first place through annotation.

Being retrievable, this information/knowledge is also reusable. Markup is also *added value* because it enhances reusability, both in the sense that it makes research easily replicable and in the sense that it makes the data readily shareable by a variety of potential users and for a range of different uses ('[. . .] the annotations themselves spark off a whole new range of uses which would not have been practicable unless the corpus had been annotated', Leech 2005, online). Not only does annotation facilitate the sharing of the corpus as resource, but it also encourages the sharing of analytic tools and of progressive results. Annotation provides a trace of the interpretative work carried out on the texts.

> Metadata plays a key role in organizing the ways in which a language corpus can be meaningfully processed. It records the interpretative framework within which the components of a corpus were selected and are to be understood.
>
> (Burnard 2004:15)

The annotation of a corpus, the selection and application of the tag set, expresses a theory about the texts: deciding which are the important characteristics that make up the identity of a source (in the case of the press, for example, the type of news—editorial, op-ed, report—the type of newspaper—tabloid vs. quality—etc.), or deciding on a unit of analysis for that source (e.g., all the articles from a single newspaper, or a single article) are operations that involve several degrees of selection, thus interpretation.

The mere fact of looking at texts in terms of uses, transforming them from texts to *textual resources* (Atkins, Levin and Zampolli 1994), implies a large amount of interpretative work, precisely because use is interpretation. More specifically, the practice of marking up is interpretation in that it involves a manipulation of the data; the text is preserved in its integrity but we superimpose on it a structure that 'speaks' of the text. The very term *markup* is borrowed from the publishing and printing business, where it indicates the instructions for the typesetter that are written on a typescript or manuscript copy by an editor and in this sense 'compilers have the responsibility typically associated with an editor' (Atkins, Levin and Zampolli 1994:34).

The annotation process is of course supported by automation, but because of the ambiguous nature of language there is a constant need for human intervention and a great amount of manual work needs to be done.

Human beings can disambiguate problems, but as humans we are prone to error and inconsistency and our choices, interpretations, and often compromises have to be specified and checked against a formalized annotation scheme in order for them to be consistent throughout the corpus.

In marking up the *CorDis Corpus*, consistency was our main concern. *CorDis* is a composite corpus, with specific problems: the variety of its text types, different levels of annotation to be managed, and different variables to be equally taken into consideration. We have dealt with differences of origin (British English, American English), differences of genre (judicial inquiry, press briefings, parliamentary debates, print and TV news), and differences of mode (writing, published official transcripts, informally transcribed speech). Originally the six different subcorpora assembled by different research groups already contained some markup on the basis of categories that the original compilers and researchers wanted to investigate but this had not been carried out according to shared norms, was not TEI-conformant, and in many cases had not been consistently applied.

Our goal consisted in consolidating all this in a single corpus, which had to be coherently marked up without losing the information which had been added by these initial attempts. In our operation of interpretation through markup we had to deal with the constraints of pre-existing interpretation, trying to preserve its richness but at the same time negotiating categories and labels. Layers of interpretation start piling up from the moment research objectives are posed, all through the process of corpus design, representation, and, of course, annotation. In addition each step involves the intervention of a number of different people. Each phase and each contribution produced knowledge about the corpus and was therefore part and parcel of the research process, but this compositeness also increased the global complexity, multiplying categories and favouring overlapping of annotation levels. It was therefore essential to strive towards some kind of standardization. 'Standardization of annotation practices can ensure that an annotated corpus can be used to its greatest potential' (Kahrel, Barnett, and Leech 1997:231), and as we will show in this chapter, aiming for TEI conformance gave the corpus a sound structure, enhancing reliability and favouring (re)usability.

1.2.3. Preliminary Information: Markup Language and Annotation Schemes

CorDis is an XML-valid, TEI-conformant corpus. Although it is not our aim here to dwell on technicalities, a few clarifications are in order to explain what these two expressions imply. The data were encoded using XML (extensible markup language), a metalanguage that enables compilers to design their own customized markup conventions for different types of documents. To say that a document (or an entire corpus) is XML-valid means first of all that it must be *well-formed*, that is, it must comply with the rules of the XML syntax. For instance, well-formed documents must

be formed of members of a set of elements, each of which may have a set of attributes, and each attribute must be assigned a value. Elements containing other elements or portions of text must be preceded by a start-tag and followed by an end-tag with the forms <elementname> and </elementname>. Attributes and their values must be stated on start-tags, taking the form <elementname attributename1="value1" attributename2= "value2" [. . .]>.

Validity, however, goes beyond well-formedness, as *valid* documents must further conform to a schema of some kind, that is, they have to be formally checked against it. The schema is, trivially speaking, a 'declaration' of what markup is allowed/required where a specification of syntax rules. However, if, as we saw in section 1.2.2, encoding consists in making interpretations of a text explicit, then the semantics of the markup should also be specified. It is here that the TEI Guidelines come into play. The guidelines, of which we have employed the P5 version (Sperberg-McQueen and Burnard 2007, online), are intended to provide standards for data interchange between researchers using different systems and applications and to suggest principles for the encoding of texts in the same format (Sperberg-McQueen and Burnard 2007, online). To be more precise,

> [t]hey provide means of representing those features of a text which need to be identified explicitly in order to facilitate processing of the text by computer programs. In particular, they specify a set of markers (or tags) which may be inserted in the electronic representation of the text, in order to mark the text structure and other textual features of interest.
>
> (Sperberg-McQueen and Burnard 2007, online)

For convenience, tags for the elements and attributes defined by the TEI Guidelines are grouped in 23 modules (cf. Sperberg-McQueen and Burnard 2007, online for the full list), each of which contains a set of declarations used to define elements/attributes and their characteristics. Each module is typically associated with a specific usage; for instance, the module *spoken* is intended for use with transcribed speech, the module *analysis* is designed to provide simple analytic mechanisms, the module *linking* caters for segmentation, alignment, and linking both within and between texts, and so on. Modules can be variously combined to form a *TEI-conformant* schema against which documents must be validated.[5] To be TEI-conformant, then, a document must be annotated using the tags that are included in the TEI modules and for which declarations are delivered in the associated schema. Moreover, each TEI-conformant text must necessarily be preceded by a TEI header, that is, an encoded unit of information containing metadata. The header provides a set of descriptions and declarations regarding the document itself and the source it is taken from (e.g., bibliographic data), its profile (categorizations of the text, and, for transcribed speech, information about the setting and the participants involved), its encoding, and its history of revisions (if any).

So far, we have generally referred to a supposedly *unitary* TEI document/text. However, a TEI text can also be *composite*, as is clearly the case with a corpus. Although the encoding of a corpus is based on the same principles as the encoding of a single text, the TEI Guidelines specifically provide for annotation of large collections of texts. Customization of markup for a multimode, multigenre corpus like *CorDis* implies using a combination of elements drawn from many different modules, and defining a corpus as a series of <TEI> documents sharing a common TEI header—the corpus header—which includes such information as bibliographic data for the corpus as a whole, the various text categorization schemes employed (what we have termed partitions and their classes), and features of encoding and revision which are shared by all the documents in the corpus. The corpus header is a separate file obtained as a modified version of the standard TEI header. It is a fundamental unit of information, in that besides containing important documentary data it also provides specific processing directives for indexing and searching applications like *Xaira*.

Xaira will be described in some detail in the following section. However, it is worth spending a few words here on the *Xaira* IndexTools Utility. This is used to construct a database which makes the corpus *Xaira*-searchable. Its main function is 'to collect information about the corpus to be supplied additional to that present in any pre-existing corpus header, and to produce a validated and extended form of the corpus header' (OUCS 2006b). Moreover, and more interestingly for the purposes of this chapter, it can be used to run the indexer and test its output. We will shortly come back to *testing*. Suffice it to say here that in addition to the corpus header, and of course the files making up the corpus proper, the indexer requires: a corpus parameter file, which defines the name (of the corpus) and the locations of the files required and to be created by the indexer, another file listing the files making up the corpus, and a bibliography file, which contains 'descriptive metadata about each source text making up the corpus' (OUCS 2006a).

1.3. MAP

Now that the methodological framework has been set, it is our aim to show the path we have constructed through the data, making reference to the practices and tools adopted. The translation of the conceptual architecture into an operative structure sprang from a series of questions concerning the harmonization process. The gradual and recursive annotation work will be illustrated through examples, highlighting some of the problems encountered and the strategies elaborated to overcome them.

1.3.1. Architecture and Tools: Going *Xaira*

We have already discussed the benefits of annotation and we have sketched the characteristics of XML TEI conformant markup, introducing the