The *CorDis* Corpus: Mark-up and Related Issues

Letizia Cirillo,[1] Anna Marchi[2] and
Marco Venuti[3]

1. Introduction

CorDis is a large, XML, TEI-conformant, POS-tagged, multimodal, multigenre corpus representing a significant portion of the political and media discourse on the 2003 Iraqi conflict. It was generated from different sub-corpora which had been assembled by various research groups, ranging from official transcripts of Parliamentary sessions, both in the US and the UK, to the transcripts of the Hutton Inquiry, from American and British newspaper coverage of the conflict to White House press briefings and to transcriptions of American and British TV news programmes. The heterogeneity of the data, the specificity of the genres and the diverse discourse analytical purposes of different groups had led to a wide range of coding strategies being employed to make textual and meta-textual information retrievable.

The main purpose of this paper is to show the process of harmonisation and integration whereby a loose collection of texts has become a stable architecture. The TEI proved a valid instrument to achieve standardisation of mark-up. The guidelines provide for a hierarchical organisation which gives the corpus a sound structure favouring replicability and enhancing the reliability of research. In discussing some examples of the problems encountered in the annotation, we will deal with issues like consistency and re-usability, and will examine the constraints imposed on data handling by specific research objectives. Examples include the choice to code the same speakers in different ways depending on the various (institutional) roles they may assume throughout the corpus, the distinction between quotations of spoken or written discourse and quotations read aloud in the course of a spoken text, and the segmentation of portions of news according to participants interaction and use of camera/voiceover.

2. The *CorDis* corpus

The *CorDis* corpus is the product of a research project called ‘Corpora and Discourse: A quantitative and qualitative linguistic analysis of political and media discourse on the conflict in Iraq in 2003’. The project, which involved various Italian Universities coordinated by the University of Siena, was aimed at developing analytical techniques based on CADS (Computer-Assisted Discourse Studies) able to “trace the progress of political messages from their inception, through their negotiation with the press to their reporting to the public, with particular attention to the linguistic-rhetorical mechanisms

[1] Dip. SITLeC, University of Bologna
e-mail: lcirillo@sslmit.unibo.it
[2] School of English, Communication & Philosophy, Cardiff University
e-mail: marchia@cardiff.ac.uk
[3] Dip. di analisi dei processi ELPT, University of Naples Federico II
e-mail: venuti@unina.it

employed" (Morley 2007). *CorDis* is a multimodal, multigenre corpus containing over five million tokens (corresponding to about 50,000 types) divided into six sub-corpora according to text type and text source: Hutton Inquiry, US parliamentary proceedings (Congressional Record), UK parliamentary proceedings (Hansard), White House press briefings, transcribed TV news programmes (from BBC and CBS), and newspaper articles from both British and American newspapers (further divided into editorials, op-eds, and reports).

As annotators we had to make sure that the corpus was marked up according to its intended purposes and complied with specific standards, specifically that it was valid XML which was TEI-conformant. We received the texts as a loose collection selected and compiled by different research groups who had already partially annotated the texts in an attempt to define categories relevant to their different research objectives. These annotations were not in XML and differed substantially from group to group. Clearly, the process of creating a unified body of codified texts posed a number of philosophical and practical issues, involving deciding how far the existing annotation should be 'standardised' and choosing appropriate tag sets. Before moving to the discussion of specific problems of annotation and detailing them by means of examples, it is worth here spending a few words on some more general aspects of the *CorDis* mark-up.

The composite nature of *CorDis* required us to draw upon different TEI modules, including the *core* tag set, a combination of *base* tag sets, and some *additional* tag sets (*cf.* Sperberg-McQueen and Burnard 2007: Introductory note). By way of simplification, we can divide the tags used into three main classes: tags containing editorial metadata, tags containing analytic metadata, and tags containing descriptive metadata (*cf.* Burnard 2004). The first class includes markers used to make interventions like omission or correction explicit (e.g. <gap desc= "omitted_from_quote"/>; the second groups structural tags and tags used to define style and textual functions (such as emphasis or quotation, e.g. <s> most critics denounced the programme as <q who="_expert"> "unduly non-intrusive" </q></s>); the third typically provides information related to the context in which the text was produced/received (e.g. <time> 10.30 am</time>). In the *CorDis* corpus most tags belong to the second class, particularly they are of the structural kind, ranging from large textual divisions, through paragraphs, utterances and sentences to words. All words were tagged by part of speech (POS) and lemma using the UCREL CLAWS7 service and the corpus was then indexed with XAIRA (XML Aware Indexing and Retrieval Architecture, developed by Lou Burnard and Tony Dodd, Oxford University Computing Services). The result of the entire process is a corpus containing twelve million XML elements altogether. If we exclude word and punctuation tags the total number of elements in text files is 511,773 corresponding to twenty-seven types of elements used in the body of texts (i.e. excluding text headers).

3. Making the case for harmonisation

In the present paper we move from the assumption that the extremely expensive and time-consuming work of annotation is worthwhile insofar as mark-up provides *added value* (*cf.* Leech, 1997a; section 5 below).[4] Annotation makes it possible to sharpen the analysis, making built-in information retrievable and allowing the user to access

[4] We will here make no distinction between 'mark-up' and 'annotation' (*cf.* McEnery, Xiao and Tono 2006: 29), using the two terms interchangeably to refer to both contextual/structural metadata (Burnard, 2004) and "interpretative linguistic information" (Leech 1997a: 2).

knowledge about the data in the corpus that would be lost if it were not made explicit through annotation.

In the case of *CorDis* it is the specific story of the project and the very nature of the corpus that made the role of mark-up not only valuable but vital. *CorDis* is a heterogeneous collection of texts and text types that find a common core in their topic, the Iraq war of 2003, and in the research purpose, corpus-assisted discourse studies, (*cf.* Partington 2004), but in order to become a corpus it needed to be harmonised through annotation. We argue that annotation, with specific reference to the XML-valid, TEI conformant mark-up, is not merely an accessory to the corpus, but that it is its backbone, since it is what makes the corpus usable as a unified body of texts.

As already mentioned in section 1, *CorDis* is made up of different, mildly annotated sub-corpora originally assembled by different research groups in order to respond to diverse research questions. The sub-corpora and the analyses already conducted on the texts had to be preserved but consolidated in a single corpus. The final product should not lose the initial imprints and specificities but should at the same time unify the sub-corpora on the basis of shared features. Our chief concern was to make it possible to use *CorDis* as a collective resource by an efficient management of modularity. XAIRA proved to be a valuable and effective tool in this respect, enabling us to proof the mark-up and test the output.

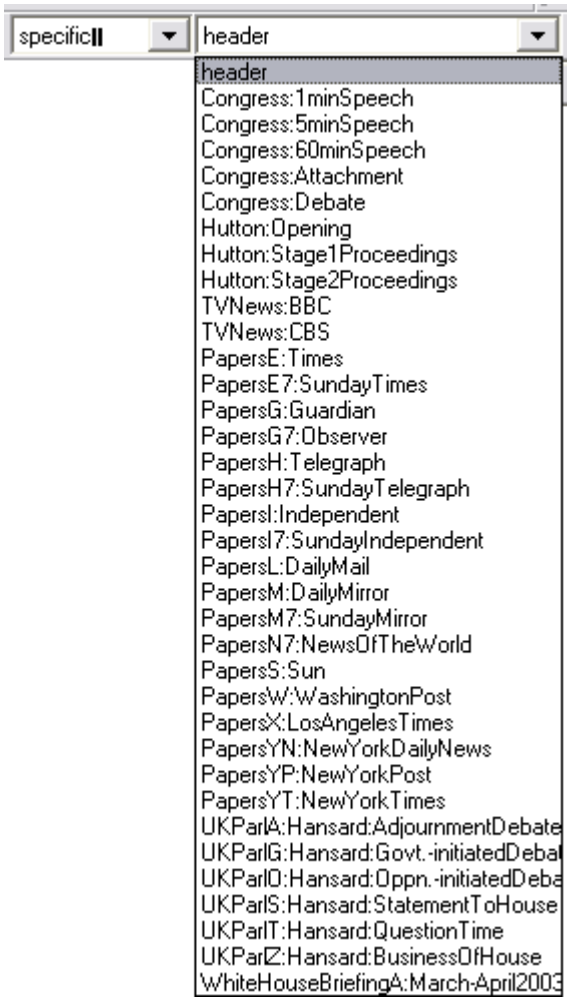

Figure 1: XAIRA's 'specific' partition for *CorDis*

In particular, XAIRA's indexing of the mark-up allows the texts to be re-divided by means of partitions that enable the user to: a) merge the original sub-corpora according to transversal parameters: *origin* (GB vs. US), *mode* (written – newspapers, spoken – TV news, and official transcripts – the Hutton Inquiry, Congressional Records, Hansard, White House Press Briefings); b) recover the original categories assembled by the various groups: *source* (Hutton Inquiry, White House Press Briefings, Congressional Records, Hansard, Newspaper editorials, Newspaper Op-eds, Newspaper reports and TV news) and *whodunnit* (indicating the groups responsible for compilation); c) split the sub-corpora into smaller units, i.e. *specific* sources or homogeneous units composing the initial sub-corpora (as exemplified in Figure 1).

The making (and the marking-up) of *CorDis* implied combining different materials as well as combining the work, the interpretation and the desiderata of a number of different people. In order to handle such a growing complexity standardisation of mark-up became an obliged goal. In this respect a few clarifications are in order. Any

annotation scheme is heavily dependent on the corpus to which it is applied, particularly on its size and intended use, which means that no annotation scheme can be taken as an absolute standard. As stated by Kahrel *et al.* (1997: 234), "[w]hile standards must be explicit and usable, they cannot be too stringent or limiting". In other words, size- and task-dependence call for a high degree of *flexibility*. For this reason we prefer to speak of *harmonisation* (*cf.* Leech, 1991: 24), rather than standardisation, our goal being an annotation system that meets both the *annotator*'s and the *user*'s requirements (*cf.* Leech 1993: 279; section 5 below).

The process of harmonisation of the *CorDis* corpus was gradual and recursive. It implied an ongoing editing work on the corpus, and progressive transformations of raw text or minimal annotation into increasingly refined XML-valid, TEI-conformant versions aimed at making mark-up "concise", "perspicuous" and "analysable" (*cf.* Leech 1997b: 25). A particularly telling example of this evolution is the coding of speaking turns in TV-news, as shown in 1) below:

1) a [Rageh Omaar] (Rageh Omaar Baghdad)
Finally the war has come to the heart of Baghdad. [...] Rageh Omaar, BBC News, Baghdad.

b <R REP>
<WZ REP>
<WZ VO> [Rageh Omaar] (Rageh Omaar Baghdad)
Finally the war has come to the heart of Baghdad. [...] Rageh Omaar, BBC News, Baghdad.
</WZ VO>
</WZ REP>
</R REP>

c <div2 type="report" resp="reporter:warzone">
<u who="Omaar_Rageh" sex="m" role="reporter:warzone" dialect="en-GB"
type="voiceover">
<writing type="subtitle"><s n="20"> Rageh Omaar Baghdad</s></writing>
<s n="21"> Finally the war has come to the heart of Baghdad.</s> [...]
<s n="37"> Rageh Omaar, BBC News, Baghdad.</s>
</u> […] </div2>

[<w> and <c> tags omitted]

Example 1: From raw text to the *CorDis* annotation scheme

Example 1 illustrates the passage from the initial transcript (a), simply signalling the speakers' names, through a basic annotation of features of interest to the group in question (b), such as the type of news (in this case a report), the person responsible for producing it (a reporter), the speaker (a war-zone correspondent), and the link between the words he utters and the screen image (in this case voiceover rather than to camera), to the final product (c), where information is made more explicit and ordered depending on whether it refers to text structural divisions (here <div2 type="report" resp="reporter:warzone"> or speakers (<u who="Omaar_Rageh" sex="m" role="reporter:warzone" dialect="en-GB" type="voiceover">).

Harmonisation has two major and intertwined objectives (and effects): consistency and reusability. *Consistency* is needed because of the intrinsic interpretative nature of annotation, and "[…] because the human analyst is susceptible to error and inconsistency, the mental interpretation of what is correct has to be sharpened and made explicit through the specification of an annotation scheme" (Baker 1997: 244). Annotation schemes make categories explicit through the selection of appropriate tag sets, preventing an uncontrolled multiplication and/or overlapping of categories, while at the same time granting categorisations a reasonable degree of delicacy (as we will show in section 4 by means of examples). Being consistently annotated, a corpus is also *re-usable*. As pointed out by Leech (2005), corpora are often exploitable for a long time

after their origin and in ways not envisaged by their originators. Moreover, "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).

In the following two sections we will further investigate the notions of consistency and re-usability. Particularly, in section 4 using some examples we will comment on the efforts made to satisfy the requests of various researchers while combining their variously assembled resources so that they may 'hang together'; in section 5 we will draw some conclusions from the observations made on the examples in section 4, using these observations to value the role of mark-up in sharing and encouraging research.

4. Striving for consistency

Marking up *CorDis* was not an easy task. As annotators we had to consolidate the various sub-corpora into a single corpus, but we were also required to preserve all necessary and relevant information regarding text types, genres and interaction formats. Given the heterogeneity of *CorDis*, this posed a huge challenge in terms of consistency of annotation, as we were fully aware that any solution we would find for specific requests related to a specific sub-corpus may well not suit the other sub-corpora or meet the needs of all research groups. In the present section we will consider some examples to show how we coped with global as well as particular annotation problems.

The fact that all texts in the corpus are centred on the same topic (i.e. the war in Iraq of 2003) attributes a highly intertextual nature to *CorDis* and explains why many of the key players involved in the Iraqi conflict appear in more than one sub-corpus. Given the specificity of each sub-corpus, however, some of these characters are attached varying institutional and/or situational roles. For instance, the former British Prime Minister, Tony Blair, appears in all sub-corpora (except the Congressional Record and the White House press briefings) either as a participant in the interaction or as a quoted source in the media. As a participant in officially transcribed interactions his role is inextricably linked to the institutional context in which the interaction takes place. Thus he is a member of the Labour Government in the UK Parliamentary proceedings, as illustrated in 2); he is coded as 'witness' (as opposed to 'judge' and 'counsel') in the Hutton Inquiry, where, in addition, due to the highly formal character of the setting, he is referred to with his full name, i.e. Anthony Charles Lynton Blair as in 3); and he takes on the role of 'legitimated person' in the TV news sub-corpus, as shown in 4). In particular, this latter choice was motivated by the peculiar interests of the TV news research group, who wanted to identify utterances spoken by journalists (further classified according to their specific journalistic activity as newsreader, reporter, correspondent, war-zone correspondent, and embedded reporter; *cf*. example 1) in section 3) as opposed to utterances spoken by non journalists, and demanded that non journalists were further classified in 'legitimated person', 'military' or 'person in the street' on the basis of their authoritativeness.

2) <u who="Blair_Tony" sex="m" role="Labour:Gov" dialect="en-GB"> <s n="244"> On the nature of the threat, it is the UN Resolution that described Saddam Hussein's programme of weapons of mass destruction as a threat, so that was established by the international community. </s> […] </u>

[<w> and <c> tags omitted]

Example 2: Attributes and attribute values of the <u> element in 'Hansard'.

3) <u who="Dingemans_James" sex="m" role="counsel" dialect="en-GB" type="question"> <s n="182"> Is there anything from your statement to Parliament that you wanted to emphasise? </s></u> <u who="Blair_AnthonyCharlesLynton" sex="m" role="witness" dialect="en-GB" type="response"> <s n="183"> I think the only thing, as I do in my witness statement to you, is just to emphasise the fact that I make it clear what I perceived the threat to be. </s> […] </u>

[<w> and <c> tags omitted]

Example 3: Attributes and attribute values of the <u> element in 'Hutton Inquiry'.

4) <u who="Blair_Tony" sex="m" role="legPerson" dialect="en-GB" type="camera"> <s n="206"> On Tuesday night I gave the order for British forces to take part in military action in Iraq. </s> […] </u>

[<w> and <c> tags omitted]

Example 4: Attributes and attribute values of the <u> element in 'BBC'.

Examples 2)-4) illustrate how the use of the attribute 'role' within the <u> element makes it possible to retrieve, whenever necessary and using dedicated applications (e.g. XAIRA), all the utterances by any person with a given role. For instance, by searching for values of the 'role' attribute matching 'Labour:Gov' all utterances by participants encoded as members of the Labour Government will be found. The same applies to other attributes and their corresponding values. Clearly, multiple searches are possible by combining two or more attributes (e.g. searching for utterances spoken by female members of the Labour Government, or by male Conservative Backbenchers, *etc.*).[5] Similar searches can be performed using the 'type' attribute, which is assigned to the <u> element in the Hutton Inquiry and the TV news sub-corpora. This attribute is used to distinguish questions from responses or other kinds of interaction in the former, and to indicate whether an utterance is spoken with the speaker shot on camera or the voice of the speaker can be heard commenting the accompanying images in the latter (cf. examples 3) and 1) respectively).

One of the problems that we had to face from the very beginning was the tagging of quoted sources. A cursory glance at *CorDis* is enough to realize that quotations represent a global issue, as they are used extensively throughout the corpus. Two possibilities were chosen: <q> and <writing>. The element <q> with the attribute 'who' and its corresponding values allow the user to single out all spoken and written quotations attributed to a specific speaker, as in example 5):

5) <u who="Spelman_Caroline" sex="f" role="Conservative:Opp" dialect="en-GB"> […] <s n="546"> If we do not get a resolution soon, what does the Secretary of State believe will be the legal position of our troops in Iraq?</s><s n="547"> Does she stand by her statement of 26 March that the coalition has no authority <q who="Short_Clare" sex="f" role="Labour:Gov" dialect= "en-GB"> to reorganise institutions or establish a new Government</q> ?</s> […] </u>

[<w> and <c> tags omitted]

Example 5: Use of the <q> element in *CorDis*.

In the case of "a passage of written text revealed to participants in the course of a spoken text" (Sperberg-McQueen and Burnard, 2007: Appendix B) the <writing> element is used instead, with the attributes <who> referring to the person reading aloud.[6] Moreover, the 'type' and/or 'script' attributes are used to specify the type and

[5] See Table 1 in the Appendix for a list of the most common 'roles' encoded for the <u> element. The attributes of the other elements discussed in this section are listed in Tables 2–5.

[6] The <writing> element is used with a slightly different function in the TV news sub-corpus. Since the TV news group aimed at a mark-up that could account for the multimodal character of the information presented in TV news programmes (as shown with the distinction between 'camera' and 'voiceover'), words that appeared on the screen, but which were not necessarily spoken by anyone,

source of the material read aloud (e.g. newspaper articles, UN Resolutions, official records and the like), and in some cases the 'n' attribute provides the official record for the quoted text, making it possible to univocally identify its source (*cf.* example 6). The attributes 'who', 'sex', 'role' and 'dialect' refer to the person reading out the quoted text. The use of the <writing> element is particularly relevant to the researchers studying Parliamentary discourse, who were interested in finding a way of excluding portions of the text which were 'read' rather than 'spoken', in order to investigate the style and rhetoric of specific speakers on the one hand, and to identify what kind of texts these speakers use to back up their claims on the other.

6) <s n="125"> The text of H Con Res 104 is as follows: <writing who="Simpson_MichaelK.Spt" sex= "m" role= "Republican:SpeakerPt" dialect= "en-US" script="resolution" n="H Con Res 104"> <s n="126"> Whereas the United States Armed Forces, a total force comprised of active, National Guard, and Reserve personnel, are now undertaking courageous and determined operations against the forces of Saddam Hussein's regime; </s> </writing>

[<w> and <c> tags omitted]

Example 6: Use of the <writing> element in *CorDis*.

A mark-up string that is typically associated with Parliamentary discourse in *CorDis* is the so-called 'referring string' or <rs>. Members of the British Parliament and the American Congress are conventionally addressed or referred to indirectly by means of association with their constituency; typical expressions include 'the gentleman/gentlewoman from [name of US state]' in the Congressional Record and 'the/my (right) Hon (Friend the) Member for [name of constituency]' in Hansard. Such references were made explicit using the attribute 'type' of the <rs> element and indicating as attribute value the name of the corresponding speaker in the chosen format 'Surname_Name', as shown in 7) below:

7) a My Hon Friend the Member for Meriden , the shadow Secretary of State […] has repeatedly called on the Secretary of State […]

b <rs type="Spelman_Caroline" sex="f" role="Conservative:Opp" dialect="en-GB"> My Hon Friend the Member for Meriden , the shadow Secretary of State </rs> […] has repeatedly called on the Secretary of State […]

[<w> and <c> tags omitted]

Example 7: Use of the <rs> element in *CorDis*.

The examples provided so far are essentially taken from the 'spoken' and 'official transcript' portions of the *CorDis* corpus, which in fact present the largest variety of tags (editorial, analytic, and descriptive; *cf.* section 2) and the highest number of annotation problems. Nevertheless, the newspaper sub-corpus also present specific features mirroring the needs of the corresponding research group. For instance, we were requested to identify the first and last paragraph in each article. Arguably, lead paragraphs have a specific function in journalism (*cf.* White 1997), therefore they were tagged by adding an 'n' attribute with a 'first' value to the <p> element, as shown in 8) below, which also illustrates *CorDis* tagging of headlines:

---

were preserved as part of the specific genre of TV news programmes and tagged as <writing type="subtitle"> (*cf.* example 1c).

8) <div1 type="news">
<head type="main"> <s n="1"> Hussein's Baghdad Falls as Tanks Roll Through City </s>
</head>
<head type="sub"> <s n="2"> Jubilant Iraqis Take to Streets , Topple Baghdad Statue of Dictator </s> </head>
<byline> <s n="3"> By Thomas W Lippman Washington Post Staff Writer </s> </byline>
 <s n="4"> After three weeks of war , Saddam Hussein no longer rules Baghdad . </s>

[<w> and <c> tags omitted]

Example 8: Tagging of the first paragraph (lead) of a newspaper article from *CorDis*.

Similarly, because certain evaluative devices tend to cluster in the last paragraph of newspaper articles (*cf.* Morley 2004), the last paragraph of each article was tagged using the value 'last' for the 'n' attribute, as in 9):

9) <text> <body><div1 type="editorial"> […] <s n="13">Once it is won—as it undoubtedly will be—where does that put France. </s> <s n="14"> Deep in the merde, as the French say. </s> </div1> </body> </text>

[<w> and <c> tags omitted]

Example 9: Tagging of the last paragraph of a newspaper article from *CorDis*.

5. Looking ahead: re-usability or the added value of mark-up

The examples analysed in sections 3 and 4 bear witness to the work done on *CorDis* in an attempt to strike a balance between granularity of annotation and global consistency. As we have seen, a reasonable level of detail was sometimes necessary to bring linguistic analysis to emergence and keep track of the specific features of each text type and of the relevant differences between genres. On the other hand, for the sake of comparability similar features/phenomena in the various sub-corpora had to be treated similarly, meaning that mark-up had to be often general rather than specific. These two contrasting needs required us to adopt an approach that favours flexibility rather than compliance with rigid standards, and an annotation scheme able to adjust to different research interests and changing research hypotheses.

The argument for flexibility is strongly made by Ide and Brew (2000: 5), who claim that "[w]hile we cannot predict future research needs, we can predict that there will be such needs", and that therefore the "corpus architecture must be such that it can be adapted to new situations as new research paradigms emerge" (*ibid.*). In the case of *CorDis* the need for flexibility prompted two main decisions, namely the use of TEI and the selection of tags codifying mainly the structural organisation of texts. The TEI model of annotation was chosen because it has proved able to cater for customisation of mark-up for large multimodal and multigenre corpora like *CorDis*. Specifically, TEI enabled us to create our own annotation scheme by combining elements taken from many different modules. As to the choice of tags, to avoid misunderstandings over their meaning and inconsistencies in their application, we have tried to rely as far as possible on "consensual categories" (Leech 1997a: 7). Therefore, we have mainly used structural tags, as we felt they were the most widely accepted and understood.

Overall, this paper has shown that the practical usefulness of mark-up goes well beyond time- and money-related issues. We firmly believe that mark-up, provided it complies with the requirement of consistency and flexibility discussed above, enhances comparability of data, comparability being, at least in our case, both internal – among sub-corpora – and external – with other corpora. The examples considered show that annotation is a way of bringing interpretative work to emergence and providing a record

of past analyses. Hence, not only does it facilitate the sharing of a corpus as resource, but it also encourages the sharing of analytic tools and of progressive results, ultimately favouring re-usability of data and replicability of research.

References

Baker, J.P. (1997) Consistency and Accuracy in Correcting Automatically Tagged Data, in R. Garside, G. Leech and T. McEnery (eds) Corpus Annotation – Linguistic Information from Computer Text Corpora, pp. 243–50. London/New York: Longman.

Burnard, L. (2004) Metadata for Corpus Work. Available on-line from http://users.ox.ac.uk/~lou/wip/metadata.htm (accessed: 23 June 2007).

Ide, N. and C. Brew (2000) Requirements, Tools, and Architectures for Annotated Corpora, in Proceedings of the EAGLES/ISLE Workshop on Meta-Descriptions and Annotation Schemas for Multimodal/Multimedia Language Resources and Data Architectures and Software Support for Large Corpora, pp. 1–6. Paris: European Language Resources Association.

Kahrel, P., R. Barnett and G. Leech (1997) Towards Cross-linguistic Standards or Guidelines for the Annotation of Corpora, in R. Garside, G. Leech and T. McEnery (eds) Corpus Annotation – Linguistic Information from Computer Text Corpora, pp. 231–42. London/New York: Longman.

Leech, G. (1991) The State of the Art in Corpus Linguistics, in K. Aijmer and B. Altemberg (eds) English Corpus Linguistics: Studies in Honour of Jan Svartvik, pp. 8–29. London/New York: Longman.

Leech, G. (1993) 'Corpus Annotation Schemes'. *Literary and Linguistic Computing* 8: 4, 376–81.

Leech, G. (1997a) Introducing Corpus Annotation, R. Garside, G. Leech and T. McEnery (eds) Corpus Annotation – Linguistic Information from Computer Text Corpora, pp. 1–18. London/New York: Longman.

Leech, G. (1997b) Grammatical Tagging, in R. Garside, G. Leech and T. McEnery (eds) Corpus Annotation – Linguistic Information from Computer Text Corpora, pp. 19–33. London/New York: Longman.

Leech, G. (2005) Adding Linguistic Annotation, in M. Wynne (ed.) Developing Linguistic Corpora: A Guide to Good Practice. Oxford: Oxbow Books. Available on-line from http://ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm (accessed: 23 June 2007).

McEnery, T., R. Xiao and Y. Tono (2006) Corpus Based Language Studies. London/New York: Routledge.

Morley, J. (2004) The Sting in the Tail. Persuasion in English Editorial Discourse, in A. Partington, J. Morley and L. Haarman (eds) Corpora and Discourse, pp. 239–55. Bern: Peter Lang.

Morley, J. (2007) Introduction – The History of CorDis. Paper presented at the CorDis Colloquium in Pontignano, 26 January 2007.

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

Sperberg-McQueen, C. M. and L. Burnard (eds) (2007) TEI P5: Guidelines for Electronic Text Encoding and Interchange. Available on-line from

http://www.tei-c.org/release/doc/tei-p5-doc/html/ST.html (accessed: 23 June 2007).

White, P.R.R. (1997) Death, Disruption and the Moral Order: The Narrative Impulse in Mass-Media Hard News Reporting, in F. Christie and J.R. Martin (eds) Genres and Institutions: Social Processes in the Workplace and School, pp. 101–133. London: Cassell.

Appendix

witness
counsel
journalist
podium
Labour:Gov
Democrat
judge
Labour:Bb
Republican
Conservative:Opp
newsreader
Conservative:Bb
Republican:SpeakerPt
correspondent
military
reporter:warzone
reporter:embed
legPerson
Republican:Chairman
person-in-street

Table 1: List of the twenty most frequent values for the 'role' attribute of the <u> element.

Dingemans_James
_journalist
McClellan_Scott
Knox_Peter
Hutton_Brian
Fleischer_Ari
Gompertz_Jeremy
McLeodScarlett_John
Gilligan_Andrew
Hatfield_RichardPaul
Tebitt_KevinReginald
Straw_Jack
Blair_Tony
Wells_Brian
Campbell_AlastairJohn
Teare_Pamela
Howard_Martin
Sumption_Jonathan
Sambrook_Richard
Caldecott_Andrew

Table 2: List of the twenty most frequent values for the 'who' attribute of the <u> element.

question
response
camera
other_interaction
voiceover
telephone
headline
videophone

Table 3: List of the values for the 'type' attribute of the <u> element.

type
who
sex
role
dialect
script
n

Table 4: List of the attributes of the <writing> element.

type
sex
role
dialect

Table 5: List of the attributes of the <rs> element.