An analytic framework for the validation of language corpora

Work Package 2
Report for the ELRA Corpus Validation Group

Paul Baker; Lou Burnard; Anthony McEnery; Andrew Wilson

1 Dec 1997

1 Introduction

A number of approaches might be taken for the establishment of an appropriate analytic framework within which the validation of language corpora might be carried out. One might, for example, decide on *a priori* grounds that validation of a multi-purpose resource such as a language corpus could only be performed with respect to a particular application: in such a case, one might need to define a different analytic framework for each corpus/application pair. The approach taken here however has been to determine *empirically* that set of textual features which current corpus users appear to agree should be captured (encoded) — whether to maximize the re-usability of the resource, or for other reasons.

This document describes how we set about gathering such evidence, and what the results of our analysis indicated. In a subsequent deliverable, we will assess the implications of these results for the automation of appropriate validation criteria for language corpora.

Data was collected from three distinct sources:

Actual practice: as demonstrated by a large and varied sample of about twenty different corpora currently in use

User requirement: as identified in the results of a survey questionnaire sent to a wide variety of corpus users

External standard: as specified by relevant published international standards or guidelines.

2 The Sample Corpora

In selecting corpora to review, we attempted to include recently constructed and well-established corpora, and to sample for a variety of languages and text modes, restricting ourselves however to corpora which were likely to be readily accessible or of major interest to European corpus users. On the basis of these criteria, the following corpora were selected for review:

Table 1: Corpora examined

Corpus	*Language*	*Mode*	*Annotation*	*Date*
BNC (British National Corpus)	British English	✆ ✍	★	1994
BRO (Brown Corpus)	US English	✍	★	1960
CRA (Corpus Resources and Terminology Extraction)	English, French, Spanish	✍	★ ➢	1995
ENP (English/Norwegian Parallel Corpus)	English, Norwegian.	✍	➢	1996
HEL (Helsinki Diachronic Corpus)	Historical English	✍	.	1994

ICE (International Corpus of English)	Geographical varieties of English	✍	★	1990
LAM (Lampeter Corpus)	Historical English	✍	.	1997
LPC (Lancaster Parsed Corpus)	British English	✍	★ ❄	1991
SEC (Lancaster IBM Spoken English Corpus)	British English	✆	★	1986
LOB (Lancaster Oslo Bergen Corpus)	British English	✍	★	1960
LLC (London Lund Corpus)	British English	✆	.	1976
MUC (Message Understanding Conference)	American English	✍	.	1992
MUL (Multext)	Nine European languages	✍	★	1996
MUE (Multext East)	Six East European languages	✆ ✍	★	1997
MUS (Multext Sweden)	Swedish	✆ ✍	★	1997
PAR (Parole)	European languages	✍	★	1997
PEN (Penn Treebank)	American English	✍	★ ❄	1995
SPC (Speech Presentation Corpus)	British English	✍	★	1996
TEL (TELRI Plato Parallel Corpus)	Ten East European languages, English, Chinese	★	➢	1997
UAM Madrid Spoken Corpus	Spanish	✆	.	1992

This list gives a good range of corpora produced over the last thirty years, containing speech (✆), writing (✍), and a mixture of the two. The corpora include a wide range of European languages (Parole and Multext, for example, cover all EU official languages) and they represent work undertaken throughout Western Europe, Eastern Europe and the USA. A high proportion of these corpora were also available in an annotated form which included some form of morpho-syntactic or other analysis, indicated above by the symbols ★ (part of speech code); ➢ (aligned corpora); and ❄ (tree-banked).

For each of these corpora we reviewed a range of manuals and other documentation; we also carried out examination of the actual corpus texts in some cases. The objective was to identify the encoding practices actually adopted for each corpus, both with respect to text features and with respect to annotations. Where there was no manual to refer to, we contacted corpus builders directly. In this way, we were able to collate the information needed to develop a profile for each corpus.

To facilitate comparison amongst them, we had originally planned simply to list the union of all features marked up (actually and potentially) in all our sample corpora. However, a closer examination of available encoding standards suggested that we might do better to use one of these as the baseline for our comparisons.

For our purposes, the Corpus Encoding Standard (CES), defined by EAGLES was of most relevance. This standard defines a number of SGML document type definitions (DTDs), which are derived from the set of recommendations produced by the international Text Encoding Initiative (TEI). In examining the corpora selected, we found few (if any) textual features for which a tag was not available from this source. It therefore seemed appropriate to use this standard as the yardstick against which to compare their respective practice.

As an indication of the delicacy of the analysis carried out, we identified nearly a hundred features in all. The principal groupings tabulated were :

- Header or metadata(about 35 in all)
- primary structure (text divisions, headings, etc: 7 features)
- paragraph-level (10 features)
- miscellaneous sub-paragraph (abbreviations, dates, numbers etc.: 11 in all
- renditional features (6 distinctions listed)
- editorial features (correction, regularization etc.: 4 features)
- segmentation and linking (4 features)

The results of this cross-comparison are given in detail in Tables 2 and 3 below, and summarized in the next section.

We performed a similar cross tabulation for the subset of our sample corpora in which morpho-syntactic analysis of some kind had been applied. This analysis, given in section 11 below, demonstrates the applicability of the EAGLES recommendations for morphosyntactic analysis across a range of existing analysed corpora.

3 Findings: corpora

Clearly, we would not expect to find any use of SGML or adherence to the CES Guidelines in the earliest corpora studied here. The LOB, Brown, Helsinki and Spoken English Corpora all naturally use idiosyncratic encoding systems; what is of more interest is the high degree of overlap both between all of these early corpora, and between them and the others in terms of the features they do choose to mark up, irrespective of the particular syntactic conventions they apply. This suggests that automatically converting their conventions to TEI conformant encoding would be quite trivial, (although bringing them into conformance with the CES requirements might require the addition of a some information not readily available). Considering how widely used these corpora are at present, and have been for some time past, such a preliminary mapping at least would seem to be well worth undertaking.

Of more concern is the extent of variability in the encoding of the modern corpora. A good number of the corpora which we reviewed might reasonably be regarded as TEI conformant (BNC, CRATER, PAROLE, MULTEXT for example), many of them specifically adhering to CES Guidelines. However, others have a far less systematic approach to encoding matters. Neither the Penn Treebank nor the MUC corpora claim to conform to TEI recommendations, nor even to SGML syntactic correctness. The ICE corpus meets some of the requirements of TEI, but omits various elements required in the header, and has only recently begun to require formal SGML validation of its contributors. The TELRI corpus (or at least the 'Plato' subset of it which its designers suggested we examine) appears to be encoding different languages in different ways, with little agreement amongst its co-operating groups even about whether or not such simple features as paragraph or sentence markers should be tagged. Some groups are in a position both to articulate and to enforce validation criteria (for example, "paragraphs should be tagged using the P tag", "corpus documents should use syntactically valid SGML conforming to a specific DTD") but many apparently are not. In such a situation, corpus interchange and integration will continue to be a dispiriting uphill task.

Even those modern corpora which may be described as TEI conformant, may take different positions with respect to such issues as to whether or not a given textual feature should actually be made explicit in their encoding. For example, the BNC makes explicit in its markup the location and nature of any material (for example a picture or table) which has been omitted from a text. In the CRATER corpus (and others) such material is silently omitted, even though there may be clear reference within the text to it. A similar 'silent correction' policy is used by PAROLE.

On the specific issue of documentation, we also found great variation amongst the corpora. Gathering precise information about how particular text features have in fact been encoded in a corpus can be time consuming as well as difficult. At least for corpora which claim to be TEI-conformant, there is a readily available public description of how the encoding scheme should function, while for those which conform to CES Guidelines, there is an additional (and equally easily found) set of rules as to how the TEI scheme should be applied. With this to hand, it should be relatively easy to determine how well the corpus builder has followed the standard, particular if any deviations from it have been correctly documented, for example in the corpus header.

Turning to corpora which use their own idiosyncratic schemes, the situation is in general disappointing. Sometimes documentation takes the form of a published article, sometimes it is available on the net, and sometimes it is only available by detective work. This might be understandable for older corpora, but really cannot be excused in more recently created corpora, whose builders have had ready access to several decades experience in both the necessity for accurate contextual information or documentation and the readiest means of supplying it together with a text.

4 User Survey

To complement the study reported above, we though it would be useful also to survey the current user community by means of questionnaire. This was posted on the Lancaster website, and its presence was widely publicized during September/October 1997. Email was sent to a large number of relevant public bulletin boards and mailing lists, around the world urging interested research centres or individuals to make their views known by visiting the web site and completing a form.

The questionnaire stated that ""we are gathering feedback on what the users of corpora would like to see encoded within publically available corpora. We have devised a questionnaire which broadly outlines all of the features available in the TEI Lite markup scheme. We would like to know which of the features you would like to see being used in the encoding of corpus data. Please specify for each feature whether it is of absolute importance, or whether you would like it if possible. " and contained a series of tables, of which users were invited to complete as many as they wished. (In practice most respondents actually filled in all of them). Section 12 below lists the results for each table. Despite the length of the questionnaire, we received a total of 26 responses during the month that the survey was carried out, from corpus building centres world wide."

In addition to questions about their preferred method of delivery, and encoding system, respondents were asked to state the extent to which they would prefer corpora to mark up each of a large number of text features. A total of 137 features were grouped into 40 'header features', 42 'primary features' and 55 'morphosyntactic features'. For each of these, respondents could indicate a preference on a four point scale, valued 'essential', 'if possible', 'no opinion' and 'don't want'. Detailed results are given in Table III below, and summarised in the next section.

5 Findings of the Survey

For 'Method of delivery' and 'Encoding system', respondents' preferences were clearly stated, and may be summarised as follows:

- CD, FTP and WWW are acceptable delivery media. Diskettes and DAT tapes are not.
- SGML is greatly preferred as a mark-up language. Unicode and eight-bit character sets are preferred over seven-bit character sets

With respect to the other items analysed, the results are less clear cut. To simplify presentation of the results from this survey, we adopted the following summation procedure. In the summary tables, a feature is rated as 'mandatory' if it received a score for ' essential' greater than the three other scores combined, and as 'desirable 'if it received a combined score for' essential ' and 'if possible ' greater than its 'no opinion' and 'don't want' scores combined. In cases where this procedure resulted in a tied score (e.g. an item scoring 11 both for 'essential' and

for the other options combined), the feature was given the benefit of the doubt and the higher ranking category chosen (i.e. 'mandatory' in this case).

In addition to these summary scores, the following tables also indicate the total number of votes cast for each option, since this varied from feature to feature. The second of the two figures below indicates the number of votes cast for options other than that indicated. For example, an item rated 'mandatory' with a score of (21/4) indicates that 21 respondents out of 25 voted it essential, the remaining 4 voting for one of the other three possibilities. An item rated 'desirable' with a score of (12/9) indicates that 12 out of 21 respondents voted it either 'essential 'or ' if possible', while the remaining 9 voted it as either ' no opinion' or 'don't want'.

6 Header features

With respect to the header features, respondents rated highly only a small number of those available. The summed scores obtained were as follows:

Mandatory Source (15/7); Encoding description (13/9); Date (10/10)

Desirable Type (12/9); Publication (15/7); File description (16/6); Tag usage (11/10)

Within the Encoding description: only the Sampling description (18/4) and Project description (15/7) were rated as 'desirable', no other features being rated higher.

7 Primary Data

Within the Primary Data, only Text body (13/9) and Gap (11/11) were rated 'mandatory', while the following features were all rated as 'desirable':

- Top-level structure (14/7)
- Text divisions (19/3)
- Head elements (17/5)
- Closer elements (11/11)
- Key words (13/9)
- Paragraph (18/4)
- Spoken paragraph (15/7)
- Caption (13/9)
- Quote (15/7)
- Poem (11/11)
- List (14/8)
- Figure (13/9)
- Bibliographic citation (14/8)
- Note (15/7)
- Table (13/9)

- Abbreviation (15/7)
- Date (12/10)
- Measure (11/11)
- Name (14/8)
- Number (14/8)
- Term (11/11)
- Time (14/8)
- Foreign (15/7)
- Title (13/9)
- Bold (12/10)
- Boxed (11/11)
- Italic (11/11)
- Roman (11/11)
- Underline (11/11)
- Caps (11/11)
- Correction (11/11)
- Regularised (12/10)
- S-Unit (13/9)
- Quoted dialogue (13/9)
- Pointing (11/11)
- Reference (11/11)

No other primary data feature was seen as being of high priority.

8 Morphosyntactic and syntactic features

Few respondents considered morphosyntactic mark-up of any kind as essential. For those who did, the summation procedure outlined above gave the following results:

Mandatory Verb (13/13); Adjective (13/13); Pronoun (13/13); Adverb (13/12); Conjunction (13/12); Numeral (13/12)

Desirable Noun (13/9); Article (20/6); Adposition (16/9); Interjection (17/8); Punctuation (20/5)

These responses seem a little capricious: it is difficult to imagine why marking a numeral for example should be more important than marking a noun. A closer examination suggests that several of the items rated 'mandatory' here are only marginally so, with very close or equivalent scores to those rated 'desirable' by our procedure. We conclude that for this particular type of mark-up the choice between' essential' and 'if possible' has little significance.

As with morphosyntactic features, only a minority of respondents expressed a requirement for markup of syntactic features. Of those who did, the following features were all rated as 'desirable':

- Bracketing
- Sentence
- Clause
- Noun Phrase
- Verb Phrase
- Adjective phrase
- Adverbial phrase
- Prepositional phrase

9 Summary of Recommendations

Our survey suggests that there is a consensus amongst users of language corpora with respect to the following minimal set of recommendations.

- Corpora should be distributed by FTP, CD or WWW.
- The feasibility of corpora being held in 8 bit or UNICODE character sets should be considered carefully.
- Corpora should encode data using SGML, preferably in a TEI-conformant manner.
- Documentation of the features encoded in a corpus should be made readily available, preferably bundled together with it.
- At least the items identified above as 'mandatory ' for headers and primary data should be marked up.

For corpora which include morphosyntactic or syntactic analysis, the picture is less clear. Where such corpora are used, respondents seem to rate all features equally highly; for many it appears to be a case of 'all or nothing'

We conclude that, while individual corpora may vary in terms of the range and depth of features presented, those features which are encoded should conform to EAGLES guidelines, preferably making use of the EAGLES Intermediate Representation for morphosyntactic features for this purpose. This will facilitate automated and semi-automated validation of such mark-up against the control set of EAGLES features. This does not, however, preclude the use of different schemes, particularly where corpora are likely to be processed mainly by humans rather than by machines.

10 Corpora Results

Each row of the following tables refers to a textual feature specified by the EAGLES Corpus Encoding Standard, indicating whether or not this feature is also distinguished in the various corpora indicated by the columns. The value ✔ indicates that the feature is (claimed to be) marked up in the associated corpus; the value . that is not. Note that we did not validate these claims, other than by checking the relevant corpus documentation.

Notes (marked in square brackets) are given below. The codes used to identify each corpus were defined above, in Table 1.

Table 2: Header features

Feature	CES tag	BNC	CRA	LOB	BRO	LPC	SEC	LLC	PEN	ICE	MUL [A]	TEL [B]	LAM	ENP [C]	UAM	HEL	MUC [D]	SPC	PAR
SGML markup	.	✔	.	.	.	.	.	.	.	✔	✔	✔	✔	✔	✔	✔	✔	.	✔
8-bit characters	.	.	.	.	.	.	✔	✔	.	.	✔	.	?	✔	✔	.	.	.	?
Header	<cesHeader>	✔	✔	.	.	.	✔	.	.	✔	✔	✔	✔	✔	✔	.	.	✔	✔
Type	type=	✔	✔	.	.	.	.	.	.	✔	✔	✔	✔	✔	✔	.	.	.	✔
Creator	creator=	✔	✔	.	.	.	.	.	.	.	✔	✔	.	✔	.	.	.	.	✔
Version	version	?	✔	.	.	.	.	.	.	✔	✔	✔	.	✔	.	.	.	.	✔
Status	status=	✔	✔	.	.	.	.	.	.	.	✔	✔	.	✔	.	.	.	.	✔
Date Created	date.created=	✔	✔	.	.	.	✔	.	.	✔	✔	✔	✔	✔	✔	✔	.	✔	✔
Date Updated	date.updated=	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	✔	.	.	✔	.	✔
File Description	<fileDesc>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	✔	.	.	.	.	✔
Title	<titleStmt>	✔	✔	.	.	.	✔	.	.	✔	✔	✔	T	✔	.	✔	✔	✔	✔
Authorship (etc)	<respStmt>	✔	✔	.	.	.	.	.	.	✔ [D]	✔	✔[D]	✔	✔[D]	✔[D]	✔[D]	.	✔[D]	✔
Edition	<editionStmt>	✔	✔	.	.	.	.	.	.	.	✔	✔	.	✔	.	.	.	.	✔
Extent: words	<wordCount>	✔	✔	.	.	.	.	.	.	✔	✔	✔	✔	✔	.	.	.	.	✔
Extent: bytes	<byteCount>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	T	.	.	.	.	✔
Extent: how done	<extent> units	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	T	.	.	.	.	✔
Publication	<publicationStmt>	✔	✔	.	.	.	.	.	.	✔	✔	✔	✔	✔	.	.	.	.	✔
Source	<sourceDesc>	✔	✔	.	.	.	✔	.	.	✔	✔	✔	✔	✔	.	.	.	.	✔
Encoding Description	<encodingDesc>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	✔	.	.	.	.	✔
Project description	<projectDesc>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	.	.	.	.	.	✔
Sampling description	<samplingDecl>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	.	.	.	.	.	✔
Editorial description	<editorialDecl>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	.	.	.	.	.	✔

continued on next page

Table 2: Header features

Feature	CES tag	BNC	CRA	LOB	BRO	LPC	SEC	LLC	PEN	ICE	MUL [A]	TEL [B]	LAM	ENP [C]	UAM	HEL	MUC [D]	SPC	PAR
conformance	<conformance>	.	✔	.	.	.	.	.	.	.	✔	✔	.	.	.	.	.	.	✔
transduction	<transduction>	.	✔	.	.	.	.	.	.	.	✔	✔	.	.	.	.	.	.	✔
correction	<correction>	✔	✔	.	.	.	.	.	.	.	✔	✔	.	.	.	.	.	.	✔
quotation	<quotation>	✔	✔	.	.	.	.	.	.	.	✔	✔	.	.	.	.	.	.	✔
hyphenation	<hyphenation>	✔	✔	[F]	[F]	.	.	.	.	.	✔	✔	.	[F]	.	.	.	.	✔
segmentation	<segmentation>	✔	✔	.	.	.	.	.	.	.	✔	✔	.	.	.	.	.	.	✔
normalization	<normalization>	[I]	✔	.	.	.	.	.	.	.	✔	✔	.	.	.	.	.	.	✔
tag usage	<tagUsage>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	.	.	.	.	.	✔
reference system	<refsDecl>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	.	.	.	.	.	✔
clasification scheme	<classDecl>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	.	.	.	.	.	✔
Profile Description	<profileDesc>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	✔	.	.	.	.	✔
Creation	<creation>	✔	✔	.	.	.	.	.	.	.	✔	✔	T	✔	.	.	.	.	✔
language usage	<langUsage>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	✔	.	.	.	.	✔
writing system	<wsdUsage>	.	✔	.	.	.	.	.	.	.	✔	✔	.	T	.	.	.	.	✔
text classification	<textClass>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	✔	.	.	.	.	✔
translations	<translations>	.	✔	.	.	.	.	.	.	.	✔	✔	.	T	.	✔	.	.	✔
annotations	<annotations>	.	✔	.	.	.	.	.	.	.	✔	✔	.	T	.	.	.	.	✔
Revision description	<revisionDesc>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	✔	.	.	.	.	✔
change	<change>	✔	✔	.	.	.	.	.	.	.	✔	✔	✔	✔	.	.	.	.	✔
date	<changeDate>	[E]	✔	.	.	.	.	.	.	.	✔	✔	.	✔	.	.	.	.	✔

Table 3: Primary data

Feature	CES tag	BNC	CRA	LOB	BRO	LPC	SEC	LLC	PEN	ICE	MUL [A]	TEL [B]	LAM	ENP [C]	UAM	HEL	MUC [D]	SPC	PAR
Top-level structure	<cesCorpus>	.	✔	.	.	.	.	.	.	.	✔	✔	✔	.	.	.	.	.	✔
Text body	<body>	✔	✔	✔	.	.	.	.	.	✔ [K]	✔	✔	✔	✔	✔	.	✔[L]	.	✔
Headings	<head> [G]	✔	✔	.	✔	.	.	.	.	✔	✔	✔	✔	✔	.	.	.	✔	✔
Closings	<closer> [G]		✔	.	✔	.	.	.	.	.	✔	✔	.	T	.	.	.	.	✔
Keywords	<keywords>	✔[H]	✔	.	.	.	.	.	.	.	✔	✔	.	T	.	.	.	.	✔
dateline	<dateline>	.	✔	.	.	.	.	.	.	.	✔	✔	.	.	.	.	.	.	✔
paragraph	<p>	✔	✔	✔	.	.	.	.	.	✔	✔	[M]	✔	✔	.	.	.	✔	✔
spoken paragraph	<sp>	✔	✔	✔	.	.	.	.	.	.	✔	.	.	.	✔	.	.	.	✔
caption	<caption>	✔	✔	✔	✔	.	.	.	.	.	✔	.	.	✔	.	✔	.	.	✔
quote	<quote>	✔	✔	✔	✔	.	.	.	.	✔	✔	[M]	✔	✔	.	.	.	.	✔
poem	<poem>	✔	✔	✔	.	.	.	.	.	.	✔	.	✔	✔	.	.	.	.	✔
list	<list>	✔	✔	✔	.	.	.	.	.	.	✔	.	✔	[N]	.	.	.	.	✔
figure	<figure>	.	✔	✔	.	.	.	.	.	[O]	✔	.	✔	[N]	.	.	.	.	✔
bibliographic citation	<bibl>	✔	✔	✔	.	.	.	.	.	.	✔	.	.	✔	.	.	.	.	✔
footnote	<note>	✔	✔	.	.	.	.	.	.	✔	✔	.	✔	✔	.	✔	.	✔	✔
table	<table>	.	✔	✔	.	.	.	.	.	.	✔	.	✔	[N]	.	.	.	.	✔
abbreviation	<abbr>	.	✔	✔	✔	.	.	.	.	.	✔	.	.	.	✔	.	.	.	✔
date	<date>	✔	✔	.	.	.	.	.	.	.	✔	.	.	.	.	.	.	.	✔
measurement	<measure>	.	✔	.	.	.	.	.	.	.	✔	.	.	.	.	.	.	.	✔
proper name	<name>	✔	✔	.	.	.	.	.	.	.	✔	.	.	✔	.	.	✔[P]	.	✔
number	<num>	.	✔	.	✔	.	.	.	.	.	✔	.	.	.	.	.	✔[P]	.	✔
term	<term>	✔	✔	✔	✔	.	.	.	.	.	✔	.	.	✔	.	.	.	.	✔
time	<time>	.	✔	.	.	.	.	.	.	.	✔	.	.	.	.	.	✔[P]	.	✔

continued on next page

Table 3: Primary data

Feature	CES tag	BNC	CRA	LOB	BRO	LPC	SEC	LLC	PEN	ICE	MUL [A]	TEL [B]	LAM	ENP [C]	UAM	HEL	MUC [D]	SPC	PAR
linguistically marked	<DISTINCT>	.	✔	.	.	.	.	.	.	.	✔	.	.	✔	.	.	.	.	✔
foreign	<FOREIGN>	.	✔	✔	.	.	.	.	.	✔	✔	.	✔	✔	✔	✔	.	.	✔
mentioned	<MENTIONED>	.	✔	.	.	.	.	.	.	✔	✔	.	.	✔	.	.	.	.	✔
title	<TITLE>	✔	✔	.	.	.	.	.	.	.	✔	.	✔	✔	.	.	.	.	✔
rendition information														✔		✔			✔
bold	rend=BO	✔	✔	✔	✔	.	.	.	.	✔	✔	[M]	✔	✔	.	.	.	.	✔
boxed	rend=BX	✔	✔	.	✔[Q]	.	.	.	.	.	✔	.	✔	.	.	.	.	.	✔
italic	rend=IT	✔	✔	✔	✔[R]	.	.	.	.	✔	✔	[M]	✔	✔	.	.	.	.	✔
roman	rend=RO	✔	✔	✔	.	.	.	.	.	.	✔	.	✔	.	.	.	.	.	✔
underline	rend=UL	✔	✔	✔	✔[R]	.	.	.	.	✔	✔	.	.	.	.	.	.	.	✔
caps	rend=CA	✔	✔	✔	✔	.	.	.	.	✔	✔	.	✔	.	.	.	.	.	✔
editorial corrections		.	.	[S]	.	.	.	.	.	.	.	.	.	.	.	.	.	.	✔
correction	<CORR>	✔	✔	.	.	.	.	.	.	.	✔	.	✔	✔	.	.	.	.	✔
gap	<GAP>	✔[T]	✔	.	.	.	.	.	.	✔	✔	.	✔	✔	.	.	.	.	✔
regularized	<REG>	✔	✔	.	.	.	.	.	.	.	✔	.	.	.	.	.	.	.	✔
segmentation	<S>	✔	✔	✔	.	✔	.	.	✔	✔	✔	[M]	.	✔	.	✔	.	.	✔
quoted dialogue	<Q>	✔	✔	✔	.		.	.	.	✔	✔	.	✔	✔	✔	.	.	.	✔
empty pointer	<PTR>	✔	✔	.	.	.	.	.	.	.	✔	.	✔	.	.	.	.	.	✔
non-empty pointer	<REF>	✔	✔	.	.	.	.	.	.	✔[U]	✔	.	✔	✔	.	.	.	.	✔

A. Multext refers to Multext, Multext East and Multext Sweden

B. According to Nobert Volz, the Plato corpus is planned to be TEI-CES compliant. (TELRI).

C. ENP uses TEI rather than CES guidelines.

D. Different kinds of intellectual responsibility are not distinguished; only a name is recorded

E. Following TEI, CDIF supplies the change date information within the <CHANGE> element.

F. Treatment of hyphenation is discussed in associated documentation

G. Following TEI, CES also allows for more specialised tags here e.g. <OPENER>, <BYLINE>.

H. Following TEI, CDIF allows the <KEYWORDS> element to appear only within Headers

K. Implied by markers at beginning and end of text.

L. Implied by marker at beginning of text.

M. TELRI practice for these features varies; see further below.

N. Lists, figures, diagrams and tables are all omitted and marked with <OMIT> tags

O. Figures are marked with an <O> tag for untranscribed text.

P. Names, numbers and times are used in MUC-6 evaluation tasks.

Q. Contains a tag for 'box', but it is unclear what this means.

R. Underline and italics are marked with the same tag

S. Corrections are silently applied, except in cases of doubt, which are left unchanged and marked with a <SIC> tag

T. Gaps are marked (erroneously) with the TEI <OMIT> tag

U. Footnote references are marked with a <FR> tag

For comparative purposes, we list below the distinctions made in the various components of the TELRI Project's parallel corpus, consisting of several different translations of Plato's *Republic*. The information presented here was taken from http://solaris3.ids-mannheim.de/~norbert/nancy.html in November 1997.

Feature	BR	WA	BU	LJ	PR	MA	KA	BE	MO	RI	SO	BU	TI
paragraph	.	✔	✔	✔	.	✔	✔	.	✔	✔	✔	✔	.
sentence	.	✔	.	✔	✔	.	.	.	.	✔	.	✔	.
dialogue	✔	✔	✔	✔	✔	.	✔	.	.	✔	✔	✔	.
bold/italic	✔	.	.	.	.	.	.	✔	.	✔	.	✔	.

11 Morphosyntactic Features distinguished by various corpora

As EU standards, the EAGLES guidelines were taken as the starting point for the checklist, which might otherwise have become excessively diverse owing to different nomenclatures, etc.

Only morphosyntax and syntax are represented here, since there are insufficient examples and guidelines for other annotation types to make satisfactory recommendations about them at this point in time.

The ✔ indicates that this feature is explicitly or implicitly marked in each part of the corpus concerned. The ❖ indicates that it is implicitly present, but not marked; the ✧ indicates that the feature is marked in some parts of the corpus but not all.

Table 5: Morphosyntactic feature usage

Feature	BNC	LOB	BRO	LPC	SEC	PEN	ICE	CRA-E	CRA-S	CRA-F	MUL-F	MUL-G	PAR
Morphosyntax	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Noun	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
noun type	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
noun gender	.	.	.	.	.	.	.	.	✔	✔	✔	✔	.
noun number	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
noun case	.	✔	.	.	.	.	.	.	.	.	.	✔	.
Verb	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
verb person	✧	✧	✧	✧	✧	✧	.	✧	✔	✔	✔	✔	✧
verb gender	.	.	.	.	.	.	.	.	✔	.	.	✔	.
verb number	✧	✧	✧	✧	✧	✧	.	✧	✔	✔	✔	✔	✧
verb finiteness	❖	❖	❖	❖	❖	❖	❖	❖	✔	❖	❖	❖	❖
verb form/mood	✧	✧	✧	✧	✧	✧	✔	✧	✔	✧	✧	✧	✧
verb tense	✔	✔	✔	✔	✔	✔	✔	✔	✔	.	.	.	✔
verb voice	.	.	.	.	.	.	✔	.	.	.	.	.	.
verb status	.	.	.	.	.	.	.	.	.	✔	✔	✔	.
Adjective	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
adj. degree	✔	✔	✔	✔	✔	✔	✔	✔	✔	.	.	.	✔
adj. gender	.	.	.	.	.	.	.	.	✔	✔	✔	✔	.
adj. number	.	.	.	.	.	.	.	.	✔	✔	✔	✔	.
adj. case	.	.	.	.	.	.	.	.	.	.	.	✔	.
Pronoun/Determiner [A]	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
pro/det person	✧	✧	.	✧	✧	.	.	✧	✔	✔	✔	✔	.
pro/det gender	✧	.	.	✧	✧	.	.	✧	✔	✔	✔	✔	.
pro/det number	✧	✧	✧	✧	✧	✧	✔	✧	✔	✔	✔	✔	✧
pro/det possessive	✔	✔	✔	✔	✔	✔	✔	✔	✔	.	.	.	✔
pro/det case	✧	✧	✧	✧	✧	.	.	✧	✔	.	.	✔	.
pro/det category	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Pronoun type	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Determiner type	✔	✔	✔	✔	✔	✔	[G]	✔	✔	✔	✔	✔	✔

continued on next page

Table 5: Morphosyntactic feature usage

Feature	BNC	LOB	BRO	LPC	SEC	PEN	ICE	CRA-E	CRA-S	CRA-F	MUL-F	MUL-G	PAR
Article [B]	✔	✔	✔	✔	✔	.	✔	✔	✔	.	.	.	.
article type	❖	❖	.	❖	❖	.	✔	❖	✔	.	.	.	.
article gender	.	.	.	.	.	.	.	.	✔	.	.	.	.
article number	✔	✔	.	✔	✔	.	.	✔	✔	.	.	.	.
article case	.	.	.	.	.	.	.	.	.	.	.	.	.
Adverb	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
adv. degree	✔	✔	✔	✔	✔	✔	✔	✔	✔	.	.	.	✔
Adposition [C]	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
adpos. type	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Conjunction	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
conj. type	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Numeral	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
num. type	✔	✔	✔	✔	✔	✔	✔	✔	✔	.	.	.	✔
num. gender	.	.	.	.	.	.	.	.	✔	.	.	.	.
num. number	✔	✔	.	✔	✔	.	✔	✔	✔	.	.	.	.
num. case	.	.	.	.	.	.	.	.	.	.	.	.	.
num. function	.	.	.	.	.	.	.	.	✔	.	.	.	.
Interjection	✔	✔	✔	✔	✔	✔	✔	✔	✔	.	.	✔	✔
Unique[D]	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Residual[E]	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
resid. type	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
resid. number	✧	.	.	✧	✧	.	.	✧	.	.	.		
resid. gender	.	.	.	.	.	.	.	.	.	.	.	.	.
Punctuation [F]	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
punct. type	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔	✔
Syntax	.	.	.	✔	.	✔	✔	.	.	.	.	.	✔
Bracketing	.	.	.	✔	.	✔	?	.	.	.	.	.	✔
Sentence	.	.	.	✔	.	✔	.	.	.	.	.	.	✔
Clause	.	.	.	✔	.	.	✔	.	.	.	.	.	.
Noun phrase	.	.	.	✔	.	✔	✔	.	.	.	.	.	✔
Verb phrase	.	.	.	✔	.	✔	✔	.	.	.	.	.	✔
Adjective phrase	.	.	.	✔	.	✔	✔	.	.	.	.	.	✔
Adverbial phrase	.	.	.	✔	.	✔	✔	.	.	.	.	.	✔
Prepositional phrase	.	.	.	✔	.	✔	✔	.	.	.	.	.	✔

A normally implied from two separate features

B this word class is occasionally grouped in with determiner

C typically implied from adposition type (Preposition)

D implied from presence of various unique word types

E implied from various residual word types

F implied from various punctuation mark tags

G determiners (e.g. **this** book) are tagged as pronouns in the ICE tagset

12 Survey Results

A total of 26 respondents replied to the questionnaire, although not all respondents answered every question. The results are tabulated below.

1: Origin of respondents

Germany	7
Unknown	5
Britain	5
Japan	3
France	2
America	1
Belgium	1
Netherlands	1
Spain	1

2: Preferred method of delivery

medium	yes	no
CD	13	9
DAT tape	5	17
Diskette	5	17
FTP	19	3
WWW	14	8

3a: Preferred Markup Language

SGML	18
Any	7

3b: Preferred Character set

7 bits	3
8 bits	7
Unicode	8

4: Preferred Header Features

Category	*Essential*	*If Possible*	*No Opinion*	*Don't Want*
Type	7	5	9	0
Creator	6	10	6	0
Version	13	5	4	0
Status	7	4	11	0
Date Created	10	7	5	0
Date Updated	8	9	5	0
File Description	9	7	6	0
Title	11	6	5	0
Author	11	6	5	0
Edition	8	8	6	0
Extent: words	4	7	7	3
Extent: bytes	1	8	10	4
Extent: how done	6	7	8	1
Publication	7	8	6	1
Source	15	6	1	0
Encoding Desc.	13	5	4	0
project desc.	5	10	7	0
sampling desc.	10	8	4	0
Editorial desc.	5	6	11	0
conformance	1	7	14	0
transduction	1	7	14	0
correction	5	8	9	0
quotation	4	8	10	0
hyphenation	1	5	13	3
segmentation	8	6	8	0
normalization	6	6	10	0
Tag usage	9	3	8	2
Reference scheme	2	6	14	0
Classification scheme	3	3	15	1
Profile Desc.	3	3	16	0
creation	3	3	16	0
lang usage	6	4	11	1
WSD	1	3	18	0
text class	4	5	12	2
translations	0	8	13	0
annotations	3	5	13	0
Revision desc.	6	5	10	0
change	8	2	11	0
date	10	3	7	0

5: Preferred Primary Data Features

Category	*Essential*	*If Possible*	*No Opinion*	*Don't Want*
top-level structure	10	4	7	0
text body	13	5	4	0
text divisions	10	9	3	0
head elements	6	11	5	0
closer elements	4	7	11	0
keywords	4	9	7	2
dateline	2	8	12	0
para	6	12	3	1
spoken p	5	10	5	2
caption	5	8	8	1
quote	6	9	6	1
poem	2	9	10	1
list	6	8	7	1
figure	5	8	7	2
bibl. Citation	4	10	7	1
note (footnote)	4	11	7	0
table	4	9	8	1
abbreviation	4	11	7	0
date	4	8	9	1
list	3	5	12	1
measure	2	9	10	1
name	6	8	7	1
number	3	11	7	1
term (formulae)	3	8	11	0
time	4	10	7	1
distinct	4	5	12	1
foreign	4	11	7	0
mentioned	3	5	13	1
title	4	9	9	0
bold	4	8	7	3
boxed	4	7	8	3
italic	4	7	7	4
roman	4	7	6	5
underline	4	7	6	5
caps	4	6	6	6
correction	9	2	11	0
gap	11	2	9	0
regularized	9	3	10	0
s unit	4	9	8	1
quoted dialogue	6	7	8	1
pointing	3	8	10	1
reference	3	8	10	1

6: Preferred Morphosyntax Features

Category	*Essential*	*If Possible*	*No Opinion*	*Don't Want*
Morphosyntax	9	8	9	0
Noun	13	9	4	0
noun type	12	8	6	0
noun gender	10	8	8	0
noun number	12	7	7	0
noun case	11	8	7	0
Verb	13	8	5	0
verb person	12	8	6	0
verb gender	8	10	7	1
verb number	12	8	5	1
verb finiteness	10	9	6	1
verb form/mood	10	10	6	0
verb tense	13	8	5	0
verb voice	11	8	7	0
verb status	8	8	10	0
Adjective	13	9	4	0
adjective degree	11	9	5	1
adjective gender	9	11	5	1
adjective number	10	10	5	1
adjective case	10	9	6	1
Pronoun/Det	13	8	5	0
pronoun/det person	12	7	6	1
pronoun/det gender	9	8	8	1
pronoun/det number	11	5	9	1
pronoun/det possessive	11	6	8	1
pronoun/det case	11	6	8	1
pronoun/det category	10	8	7	1
pronoun-type	11	8	7	0
determiner-type	10	7	7	1
Article	12	8	6	0
article type	11	7	8	1
article gender	9	8	8	1
article number	10	8	6	1
article case	9	8	7	1
Adverb	13	7	5	0
adverb degree	12	6	6	1
Adposition	11	5	8	1
adposition type	9	8	7	1
Conjunction	13	6	6	0

continued on next page

Category	*Essential*	*If Possible*	*No Opinion*	*Don't Want*
conjunction type	13	5	7	1
Numeral	13	6	6	0
numeral type	11	6	6	1
numeral gender	9	7	8	1
numeral number	9	8	7	1
numeral case	9	7	8	1
numeral function	8	9	7	1
Interjection	9	8	8	0
Unique	6	4	15	0
Residual	7	3	15	0
residual type	6	3	16	0
residual number	6	3	16	0
residual gender	6	3	16	0
Punctuation	12	8	5	0
punctuation type	10	7	7	1
Syntax	11	6	8	0
Bracketing	8	11	5	1
Sentence	12	8	4	1
Clause	12	9	4	0
Noun phrase	12	8	5	0
Verb phrase	9	9	5	0
Adjective phrase	10	9	6	0
Adverbial phrase	10	10	5	0
Prep phrase	11	9	5	0

13 Bibliography

Published References

Atkins, S., Clear J. and Ostler, N. (1992). 'Corpus design criteria' *Literary and Linguistic Computing* 7:1, 1-16.

Ballester, A., Santamarìa, C. and Marcos-Marin, F. (1993). 'Transcription conventions used for the Corpus of Spoken Contemporary Spanish'. *Literary and Linguistic Computing*, 8:4, 283-92.

Burnard, L. (1993). 'The Text Encoding Initiative: a further report' C. Souter and E. Atwell (eds), *Corpus-Based Computational Linguistics*, pp. 37--45. Amsterdam: Rodopi.

Burnard, L. (1995) *Users' Reference Guide for the British National Corpus Version 1.0* Oxford University Computing Services.

Cover, R., Duncan, N. and Barnard, D.T. (1991). 'The progress of SGML (Standard Generalized Markup Language): extracts from a comprehensive bibliography' *Literary and Linguistic Computing* 6:3, 197--209.

Crowdy, S. (1994). 'Spoken corpus transcription' *Literary and Linguistic Computing*, 9:1, 25--28.

Francis, W. N. (1964) *Manual of Information to accompany A Standard Sample of Present-Day Edited American English* .Brown University.

Freis, U. (ed) (1994) *Creating and Using English Language Corpora* Amsterdam, Rodopi.

Garside R., Leech G., Varadi T. (undated) *Manual of Information for the Lancaster Parsed Corpus*Lancaster University.

Goldfarb, Charles *The SGML Handbook*. Oxford University Press, 1990

Greenbaum, S. and Yibin, N. (1996). 'About the ICE tagset' In: S. Greenbaum (ed.), *Comparing English Worldwide: The International Corpus of English*, pp. 92--109. Oxford: Clarendon Press.

Greenbaum S and Svartvik J. (1990) 'The London-Lund Corpus of Spoken English' in Jan Svartvik (ed) *The London Lund Corpus of Spoken English: Description and Research* Lund Studies in English 82. Lund University Press.

Hockey, S. and Walker, D. (1993).' Developing effective resources for research on texts: collecting texts, tagging texts, cataloguing texts, using texts, and putting texts in context' *Literary and Linguistic Computing*, 8:4, 235--42.

Ide, N. and Véronis, J. (eds). (1996). *The Text Encoding Initiative: Background and Context*Dordrecht: Kluwer.

Johansson, S. (1978) *Manual of Information to accompany The British Lancaster-Oslo/Bergen Corpus of British English* Department of English, University of Oslo.

Manual of Information to accompany the Spoken English Corpus Lancaster University.

Nelson, G. (1996). 'Markup systems'. In: S. Greenbaum (ed.), *Comparing English Worldwide: The International Corpus of English*, pp. 36--53. Oxford: Clarendon Press.

Sampson, G. (1993). 'The need for grammatical stocktaking' *Literary and Linguistic Computing* 8:4, 267--73.

Sperberg-McQueen, C.M. (1991). 'Text in the electronic age: textual study and text encoding, with examples from medieval texts.' *Literary and Linguistic Computing* 6:1, 34- -46.

Sperberg-McQueen, C.M. and Burnard, L. *Guidelines for electronic text encoding and interchange (TEI P3)* Chicago and Oxford, ACH-ALLC-ACL Text Encoding Initiative, 1994.

Svartvik J, Eeg-Olofsson M, Rofsheden O, Orestrom B, Thavenius C. (1982) *Survey of Spoken English: Report on Research 1975-1981*Liber Laromedel, Lund.

Internet References

British National Corpus (BNC) : http://info.ox.ac.uk/bnc/

CRATER : http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html

EAGLES recommendations : http://www.ilc.pi.cnr.it/EAGLES96/browse.html

English-Norwegian Parallel Corpus : http://www.hf.uib.no/iba/prosjekt

European Corpus Initiative: http://www.elsnet.org/resources/eciCorpus.html

ICE : http://www.ucl.ac.uk/english-usage/

Lampeter Corpus : http://www.tu-chemnitz.de/~ehe/real/lamplist.htm

MUC corpora : ftp://ftp.nosc.mil/pub/MUC/

Multext : http://www.lpl.univ-aix.fr/projects/multext/

Multext specifications and proposed standards : http://www.lpl.univ-aix.fr/projects/multext/MUL3.html

Penn Treebank : http://www.cis.upenn.edu/~treebank/

SGML Tutorial: Chapter two of Guidelines for Electronic Text Encoding and Interchange (TEI P3) : http://ota.ahds.ac.uk/ota/teip3sg/

SgmlQL SGML Query Language : http://www.lpl.univ-aix.fr/projects/SgmlQL/

TEI Guidelines for Electronic Text Encoding and Interchange (P3) : http://etext.virginia.edu/TEI.html

TEI Lite: An Introduction to Text Encoding for Interchange : http://ota.ahds.ac.uk/ota/teilite/

The SGML Web Page : http://www.sil.org/sgml/

Trans-European Language Resources Infrastructure : http://www.ids-mannheim.de/telri/telri.html

UAM Spanish Corpora : ftp://lola.lllf.uam.es/pub/corpus

Validation of language resources : http://www.icp.grenet.fr/ELRA/validat.html