TGC W30: Corpus Document Interchange Format v. 1.2

Lou Burnard

15 Sept 1992

Contents

1 Introduction

This document is the formal specification for the Corpus Document Interchange Format (CDIF) which is the target encoding to be used for all textual components of the British National Corpus.

CDIF is an application of ISO 8879, Standard Generalised Mark Up Language. This international standard provides, amongst other things, a method of specifying an application-independent document grammar, in terms of the elements which may appear in a document, their attributes, and the ways in which they may be legally be combined. [1] Documents encoded using CDIF are processable using any SGML-aware software.

In addition, the development of CDIF has been very strongly influenced by the proposals currently being developed by the Text Encoding Initiative (TEI). This international research project[2] has for its goal the development of a set of comprehensive Guidelines for the Encoding and Interchange of electronic texts amongst researchers. An initial report appeared in 1991[3], and a substantially revised and expanded version is due to appear in early 1993. Like CDIF, the TEI Guidelines are themselves an application of SGML. In designing CDIF, we have consciously attempted to conform to TEI recommendations, so that CDIF texts should also be amenable to any TEI-conformant software.

The elements and attributes proposed for use in CDIF are a distinct subset of those proposed by the TEI. In principle, components with the same names in both CDIF and TEI schemes should be assumed to have identical semantics, subject to any additional constraints specified below.

This document describes the structure supported by CDIF in a "top-down" manner. An alphabetic reference list of all possible CDIF elements and their details is also provided as an appendix. Please note that by no means all of the features described here will be present in every text of the corpus, nor, if present, will they necessarily be tagged. The reference list indicates for each textual element whether its tagging when present is

- required for CDIF conformance
- recommended
- optional

For further discussion of the question of CDIFconformance, see working paper TGCW27 on Acceptance Procedures.

2 Basic structure

The British National Corpus contains a large number of *texts* , some spoken and some written. A "text" may be a complete bibliographic item, or a sample from one, or a combination of different items. Each such text is prefixed by a descriptive *header* , in which are recorded details of the text's source, composition, encoding conventions etc. The combination of a "text" in this sense with its individual header makes up a valid CDIF element. The British National Corpus consists of a large number of such elements, prefixed by a corpus header. For more detail on the header, see section **??**.

[1] For more information on SGML, an awareness of which is assumed in the remainder of this document, see Herwijnen, Goldfarb, Bryan, Burnard in the bibliography

[2] For more information on the TEI, see Burnard, Hockey in the bibliography

[3] See Sperberg-McQueen in bibliography

Individual texts may be characterised by their internal organization and by their completeness. Taking the first of these first, we distinguish between texts which are *sequential* , that is, texts of which the components are intended to be read together as a unit, in the order in which they are encoded, and those which are *composite* , that is, texts of which the components are relatively independent and may be read in any order. A book is the classic example of the first kind; one example of the second is a newspaper column or feature in which several different stories are combined together; another might be a spoken "text" in which a number of small but unrelated conversations have been combined together, or a collection of artefacts such as theatre programmes or press cuttings. To some extent the distinction is of course an arbitrary one, depending only on the perceived internal coherence of the item in question. The attribute ORG is used to specify the kind of text concerned; this attribute may also be attached to subdivisions of a text, as further discussed below.

The sampling procedures and other design principles underlying the construction of the BNC are well documented elsewhere [4]and are not further discussed here, other than to note the existence of the COMPLETE attribute, which is used to indicate whether or not a text is *complete* , that is, whether the whole of the original source from which it was derived has been transcribed or only a sample. Certain elements (for example, editorial notes or front matter) may however be omitted, even from a text marked as complete.

Distinct tags are used for spoken and written texts. For a spoken "text", we use the tag <STEXT>, and for a written one the tag <TEXT>. [5]. These two elements have slightly differing content models as further discussed below, but behave identically in structural terms. That is, a <CDIF> element always consists of a <HEADER> followed by either a <TEXT> or an <STEXT> element.

In summary:

<CDIF> a single conformant CDIF text, comprising a header followed by either a written or a spoken text.

<HEADER> contains full bibliographic and descriptive information about a spoken or written text

<TEXT> an individual written text. Attributes include:

ORG specifies how the content of the text is organised. Values:

compo composite content: i.e. no claim is made about the sequence in which elements inferior to this one are to be processed, or their inter-relationships

seq sequential content: i.e. elements inferior to this are regarded as forming a logical unit, to be processed in the sequence given

COMPLETE specifies whether or not this text is complete or a sample. Values:

Y in principal, all of the original has been transcribed

N a sample of the original has been taken

<STEXT> an individual spoken text. Attributes include:

ORG specifies how the content of the text is organised. Values:

[4]See in particular the TGA Workpapers on Spoken and Written Corpus Design

[5]This is a departure from TEI-recommended practice, made in the interests of simplifying the DTD.

compo composite content: i.e. no claim is made about the sequence in which elements inferior to this one are to be processed, or their inter-relationships

seq sequential content: i.e. elements inferior to this are regarded as forming a logical unit, to be processed in the sequence given

COMPLETE specifies whether or not this text is complete or a sample. Values:

Y the full text of the original has been transcribed

N a sample of the original text has been taken

These tags are all required. The CDIF dtd begins by declaring parameter entities which are used to define various element classes and sequences used elsewhere within it (see further section **??** below). This is followed by declarations for all the other elements described in this document.

```
<<!>!-- 2:                                                    -->
<<!>!-- British National Corpus: CDIF 1.1                    -->
<<!>!-- ... declarations from section 3 (Element class       -->
<<!>!-- declarations) go here ...                            -->
<<!>!-- ... declarations from section 4 (Parameter entities) go  -->
<<!>!-- here ...                                             -->
<<!>!--  ****************************************            -->
<<!>!ELEMENT cdif          - -  (header,(text|stext))          >
<<!>!ATTLIST cdif              %global;                        >
<<!>!-- ... declarations from section 2.1 (High level structure  -->
<<!>!-- (written)) go here ...                               -->
<<!>!-- ... declarations from section 6 (Paragraph-level     -->
<<!>!-- chunks) go here ...                                  -->
<<!>!-- ... declarations from section 7.1 (Phrase level      -->
<<!>!-- elements) go here ...                                -->
<<!>!-- ... declarations from section 2.2 (High level structure  -->
<<!>!-- (spoken)) go here ...                                -->
<<!>!--  ****************************************            -->
<<!>!-- ... declarations from section  () go here ...        -->
```

2.1 Basic structure: written texts

Written texts exhibit a bewildering variety and richness of different structural forms. Some have very little organization at levels higher than the paragraphs; others may have a complex hierarchy of parts, sections, chapters etc. Novels are divided into chapters, newspapers into sections, reference works into articles and so forth. Such hierarchies are of importance for two reasons: firstly, they indicate meaningfully cohesive sections of the text; secondly, they provide a convenient means of locating individual segments within the text as a whole. Following the TEI, in CDIF we propose a single neutrally-named "division" element for all hierarchical divisions of this kind.

Written texts, of whatever kind, are hierarchically subdivided into divisions. The largest such subdivision is tagged <DIV1>; if this is subdivided in the source text, its components are tagged <DIV2>, and so on. The smallest recognized subdivision of a text is tagged <DIV4>. Structural subdivisions smaller than this, (but above paragraph level) are not supported by CDIF; if any occur, they will be "flattened". If a text has any structural subdivision at all, then at least those at the highest level (<DIV1>) must be identified; if any such element is further subdivided, it is highly recommended that its subdivisions be also tagged (as <DIV2> elements); further levels of subdivision within this (i.e. <DIV3> and <DIV4>) may also be supplied where appropriate, but are not required. The

N attribute may be used to carry any identification used within the text for a given division, for example, a chapter number, as further discussed below (see section **??**). The attribute TYPEshould be used to characterise all divisions at this hierarchic level (e.g. as 'chapter', 'part' etc). [6]The attributes ORG and COMPLETE are also available for use, with the same meaning as for the whole text.

A sequence of paragraph level elements of arbitrary length may precede the first structural subdivision at any level. A text may have no structural divisions within it at all.

If captured, any prefatory or appended matter not forming part of a text must be distinguished from it. The <FRONT> and <BACK> elements are provided for this purpose. No provision is made by CDIF for structural subdivision of these elements, other than at the paragraph level.

The <DIV0> tag is intended as an alternative top-level element, in the event that it proves convenient to group together existing <DIV1>elements into some higher level unit smaller than a text.

To summarize, within a written text, the following tags are used to identify major structural elements. Note that the attributes TYPE, ORG and COMPLETE applicable to <DIV1> elements are also available on <DIV2>, <DIV3>, and <DIV4>, but have not been repeated in this list.

<FRONT> material prefixed to but not forming part of a written text.

<DIV0> major subdivision of a written text, where this is not a div1.

<DIV1> major subdivision of a written text, e.g. chapter. Attributes include:

TYPE categorises the division in some respect, e.g. as a chapter, section etc.

ORG specifies how the content of the division is organised. Values:

compo composite content: i.e. no claim is made about the sequence in which elements inferior to this one are to be processed, or their inter-relationships

seq sequential content: i.e. elements inferior to this are regarded as forming a logical unit, to be processed in the sequence given

COMPLETE specifies whether or not this division is complete or a sample. Values:

Y the full text of the original has been transcribed

N a sample of the original text has been taken

<DIV2> further subdivision of a written text, entirely contained within a div1, e.g. section.

<DIV3> further subdivision of a written text, entirely contained within a div2, e.g. subsection.

<DIV4> Lowest possible subdivision of a written text, entirely contained within a div3, e.g. subsubsection.

<BACK> matter not forming part of the text body but added as an appendix or similar.

The formal definition of these elements is as follows:

[6]A set of values for this attribute will be defined at a later stage in the project.

```
<!-- 2.1:  High level structure (written)                    -->
<!-- w r i t t e n                                           -->
<!ELEMENT text          - -  (front?, %pSeq, (div0| div1)*,
                              back?)              +(%wEmpty|%cEmpty)>
<!ATTLIST text               %global;
          complete           (Y|N)               Y
          org                (compo|seq)         seq            >
<!ELEMENT front         - -  (%pSeq)                            >
<!ATTLIST front              %global;                           >
<!ELEMENT div0          - o  (head*, %pSeq;, div1*)             >
<!ATTLIST div0               %global;
          complete           (Y|N)               Y
          type               CDATA               #IMPLIED
          org                (compo|seq)         seq            >
<!ELEMENT div1          - o  (head*, %pSeq;, div2*)             >
<!ATTLIST div1               %global;
          complete           (Y|N)               Y
          type               CDATA               #IMPLIED
          org                (compo|seq)         seq            >
<!ELEMENT div2          - o  (head*, %pSeq;, div3*)             >
<!ATTLIST div2               %global;
          complete           (Y|N)               Y
          type               CDATA               #IMPLIED
          org                (compo|seq)         seq            >
<!ELEMENT div3          - o  (head*, %pSeq;, div4* )            >
<!ATTLIST div3               %global;
          complete           (Y|N)               Y
          type               CDATA               #IMPLIED
          org                (compo|seq)         seq            >
<!ELEMENT div4          - o  (head*, %pSeq; )                   >
<!ATTLIST div4               %global;
          complete           (Y|N)               Y
          type               CDATA               #IMPLIED
          org                (compo|seq)         seq            >
<!ELEMENT back          - -  (%seq)                             >
<!ATTLIST back               %global;                           >
```

2.2 Basic structure: spoken texts

Spoken texts are organized quite differently from written texts. In particular, a complex hierarchy of divisions and subdivisions seems inappropriate. However, for convenience of recording, spoken texts are subdivided into *conversations* , which thus constitute a loosely-defined structural unit. If more than one such unit is defined within a particular spoken text, each is regarded as forming a distinct <DIV> element. These may not be further subdivided.

To handle overlapping utterances, TEI and CDIF use a particular device known as an *alignment map* , discussed in section **??**below. A single alignment map, represented by the <ALIGN>element may be defined for a whole spoken text, or, more usefully, for each division of it.

The following high level structural elements may thus appear within spoken texts:

<STEXT> an individual spoken text.

<ALIGN> defines an alignment map used to synchronise points within a spoken text.

<DIV> any arbitrary division of the utterances (etc.) making up a spoken text.

The formal declaration for a spoken text thus has the following form:

```
<!-- 2.2:  High level structure (spoken)                    -->
<!-- s p o k e n                                            -->
<!ELEMENT stext         - -  (align?, %spSeq, div*)
                                              +(%sEmpty;|%cEmpty;)>
<!ATTLIST stext              %global;
          complete           (Y|N)              Y
          org                (compo|seq)        seq          >
<!-- The alignment element is defined below                 -->
<!ELEMENT div           - o  (align?, %spSeq)                >
<!ATTLIST div                %global;
          type               CDATA              #IMPLIED     >
```

3 Element classes

As well as the basic structural elements discussed so far, the CDIF scheme allows for a wide range of textual features to be tagged. These features could be classified in a number of different ways — by the kinds of texts in which they appear, their function, their optionality, or their structural similarity, or a combination of these. The current CDIF dtd classifies elements primarily by their structural similarity, as described in this section.

Three basic classes are defined:

empty Empty elements have no content. They are used to mark a point within a text for some purpose. Examples include <PB>, to mark a page break; <PTR> to mark a synchronization point in a spoken text and <DEL> to mark where some text has been omitted. They can appear anywhere within a text, spoken or written, and are therefore defined as inclusion exceptions.

phrases phrase-level elements contain only character data, or other phrase-like elements. Examples include <HI> for typographically highlighted phrases, <DATE> for dates, <PROPNAME> for proper names etc. in written texts; Phrase elements, with a few exceptions, may appear in a text only when they are contained by a higher level structural element, a *chunk*.

chunks Chunks are elements which may contain other chunks, or a sequence of phrase elements. Examples include lists or paragraphs in written texts.

A few elements are not assigned to any particular class in the CDIF scheme. These include the structural elements already discussed, and a few others.

Each of these classes is further subdivided into

- elements occurring only in spoken texts
- elements occurring only in written texts
- elements which may occur in either spoken or written texts

The CDIF DTD contains *parameter entity declaration*s for most of the above element classes. This simplifies the process of modifying the DTD, both when adding and when removing existing element declarations. Their formal declarations are as follows:

```
<!-- 3:  Element class declarations                         -->
<!-- element classes common to all texts                    -->
<!ENTITY % cEmpty 'cEmpty del' >
```

```
<!ENTITY % cPhrase 'cPhrase abbrev | add | date | label | loc | distinct
                | propName | reg | salute | sic | title' >
<!ENTITY % cChunk 'cChunk note | poem' >
<!-- element classes for written texts                          -->
<!ENTITY % wPhrase 'wPhrase hi | stage' >
<!ENTITY % wEmpty 'wEmpty lb | pb' >
<!ENTITY % wChunk 'wChunk caption | citn | list | quote | sp' >
<!-- element classes for spoken texts                           -->
<!ENTITY % sEmpty 'sEmpty event | pause | ptr | shift | unclear | vocal' >
<!ENTITY % sPhrase 'sPhrase trunc' >
<!-- there is no class  sChunk -->
```

4 Element sequences

One advantage of classifying the various elements which can appear in CDIF texts as described above is that it becomes possible to define the contents of each element at a high level, which in turn gives greater flexibility. To this end, additional parameter entities are used within the dtd to denote element sequences. The following sequences are defined:

phrase sequence a sequence containing PCDATA and written phrase level elements only: this defines the content of all written phrase level elements

spoken phrase sequence a sequence containing PCDATA and spoken phrase level elements only: this defines the content of all spoken phrase level elements

paragraph sequence a sequence containing only paragraphs or other *chunks*: this is used within <TEXT>, <FRONT>, <BACK>, <QUOTE>, and all <DIV0>, <DIV1> etc. elements.

generic sequence a sequence containing any mixture of phrase or paragraph sequences: this defines the content of all other written elements

No special name is defined for sequences of empty elements. These may appear at any point within a text. They are defined as inclusion exceptions at the highest level (<TEXT> or <STEXT> as appropriate).

The following parameter entities are used to define sequences. Their use in controlled modification of the dtd so as to check segmentation of the text is further discussed in section **??**.

```
<!-- 4:  Parameter entities                                      -->
<!-- Parameter entities for common sequences                     -->
<!ENTITY % phSeq '(#PCDATA | %cPhrase; | %wPhrase;)*' >
<!ENTITY % spSeq  '(#PCDATA | %sPhrase;)*' >
<!ENTITY % pSeq '(%cChunk; | %wChunk; | p)*' >
<!ENTITY % seq
    '(#PCDATA | %cPhrase; | %wPhrase; | %cChunk; | %wChunk;)* ' >
```

5 Attribute classes

The set of attributes which can be specified for every element in the CDIFscheme is also regarded as forming a class, in this case an attribute class, called *global*. The members of this class are as follows:

ID system-generated identifier of an item, unique within the corpus

N any name or identifier for an element, not necessarily unique within the corpus

R specifies the rendition or appearance of an element.

The class itself is represented in the DTD by a parameter entity with the following formal definition:

```
<!-- 5:  Parameter entities (cont'd)                          -->
<!-- (continuation of sec. 4)                                 -->
<!ENTITY % global '
          r                  CDATA              #IMPLIED
          id                 ID                 #IMPLIED
          n                  CDATA              #IMPLIED       ' >
```

6 Paragraph-level elements and chunks

Written texts may be organized into structural units larger than any of those classed as phrases, but smaller than any of the divisions discussed in section **??** above. For spoken texts, it is arguable that individual *utterances*or speaker turns are analogous. We have however chosen to treat them as phrase-level elements; see further section **??** below. In written texts, the most commonly found such element is the *paragraph* , but there are several others, most of which are classed as *chunks*. Their common identifying feature is that they may appear directly within structural divisions (that is, not nested within some other element). Some of them may appear in spoken texts as well as in written ones.

A list follows:

<P> a paragraph in a written text.

<HEAD> a title or heading prefixed to some division of a written text or to a poem.

<CAPTION> (1) a heading, title etc. attached to a picture or diagram, usually with deictic content (2) a 'pull quote' or other text about or extracted from a text and superimposed upon it to draw attention to it.

<QUOTE> a quotation from some author other than that of the surrounding text, usually either embedded or displayed.

<SP> contains material marked as "written to spoken", usually by the presence of a speaker prefix, for example in a play script or printed interview.

<POEM> a poem, or an extract from one, embedded or quoted within a spoken or written text.

<LIST> a collection of distinct items flagged as such by special layout in written texts, often functioning as a single syntactic unit.

<CITN> a loosely-structured bibliographic citation.

<NOTE> a foot-, end- or side-note in a written text, not forming part of the main text; any form of additional comment or gloss in either written or spoken texts.

The paragraph is formally defined as follows. It differs from the other chunk elements described in this section only in that it may not contain other paragraphs, and hence does not need to have an explicit end tag.

```
<!-- 6:  Paragraph-level chunks                                  -->
<!ELEMENT p              - o  (%seq)                              >
<!ATTLIST p                   %global;                            >
```

Each of the others, and their constituent components, is briefly described in the remainder of this section.

6.1 Headings and captions

Headings and captions serve a variety of functions in written texts. CDIF distinguishes between <HEAD> elements, which can appear only at the start of a text division and are logically associated with it (for example, chapter titles, newspaper headlines etc.) and <CAPTION> elements which are logically independent of the position they may have within a textual division (for example, captions attached to pictures or figures, "pull-quotes" embedded within the text, "by-lines" identifying authorship and provenance of a newspaper or periodical article.

One or more <HEAD> elements may appear in sequence at the start of any <DIV0>, <DIV1> (etc.) element, or at the start of a <LIST> or <POEM>. Any number of <CAPTION> elements may appear at any point within the text.

The TYPE attribute may be used to distinguish more exactly the function of the caption or heading, as indicated below. Lengthy discursive headings, often found in periodicals, should not be tagged using <HEAD>; if they are clearly the work of a sub-editor (rather than that of the author of the piece), they should be tagged as <NOTE>s; if they are extracts from the text used as pull-quotes, they should be treated as <CAPTION>s. True headings may contain only phrase-level elements, in any combination, and must appear at the start of a division. They may not contain paragraphs or other chunks.

Captions, by contrast, may appear anywhere within or between paragraph-level items. They may contain any combination of other chunks and phrase level elements, or paragraphs.

In summary,

<HEAD> a title or heading prefixed to some division of a written text or to a poem. Attributes include:

TYPE characterizes the heading in some respect. Values:

byline heading containing authorship or provenance of an article in a periodical

main a main heading (only one allowed per div)

sub a secondary heading (may be zero or more per div)

unspec not specified or unknown

<CAPTION> (1) a heading, title etc. attached to a picture or diagram, usually with deictic content (2) a 'pull quote' or other text about or extracted from a text and superimposed upon it to draw attention to it. Attributes include:

TYPE categorises the caption. Values:

byline caption containing authorship or provenance of an article in a periodical

display extra-textual caption such as a pull quote or displayed box

attached caption describing a non-transcribed item such as a figure or photograph

unspec not specified or unknown

Formal definitions for these elements are as follows:

```
<!-- 6.1:  Paragraph-level chunks (cont'd)                    -->
<!-- (continuation of sec. 6)                                 -->
<!ELEMENT head          - o  (%phSeq)                          >
<!ATTLIST head               %global;
          type               (main|sub|byline|unspec)
                                                 unspec       >
<!ELEMENT caption       - -  (%seq)                            >
<!ATTLIST caption            %global;
          type               (attached|display|byline|unspec)
                                                 unspec       >
```

6.2 Quotations

The element <QUOTE> is used to mark quotations in written texts only. A quotation is an extract from some other work than the text itself which is embedded within it, for example as an epigraph or illustration. It contains any combination of other chunks (for example paragraphs, poems, lists) but may not directly contain phrase level elements. Any reference for the citation should also be contained within it, tagged with the <CITN> or <TITLE>element. In summary:

<QUOTE> a quotation from some author other than that of the surrounding text, usually either embedded or displayed. Attributes include:

TYPE categorises the quotation in some way. Values:

inline inline quotation

display displayed and possibly indented

unspec unspecified

The <QUOTE> element is formally defined as follows:

```
<!-- 6.2:  Paragraph-level chunks (cont'd)                    -->
<!-- (continuation of sec. 6)                                 -->
<!ELEMENT quote         - -  (%pSeq)                           >
<!ATTLIST quote              %global;
          type               (inline|display|unspec)
                                                 unspec       >
```

6.3 Spoken paragraphs

The element <SP> is used to mark parts of a written text which are intended to be spoken, for example the speeches in a dramatic text or a published interview. Such parts are generally readily identifiable by the use of such conventions as speaker prefixes (the label supplying the name of the speaker) and stage directions, for which specific tags are also defined.

This element differs from the element <U> in that the latter is used *only* for speaker turns identified in a spoken text,. i.e. one which has been transcribed from audio tape. The <SP> element is used only for speaker turns identified as such in a written text.

It may begin with one or more optional <SPKR> elements identifying the speaker or speakers, followed by any combination of other chunks (for example paragraphs, poems, lists) but may not directly contain phrase level elements.

The WHO attribute is used in the same way as with the <U> attribute as a means of associating a speech with information about its speaker. It is not anticipated that this mechanism will be extensively used for written texts, but it is provided for consistency.

In summary:

<SP> contains material marked as "written to spoken", usually by the presence of a speaker prefix, for example in a play script or printed interview. Attributes include:

WHO may be used to standardize speaker identification.

<SPKR> contains the speech prefix used in the original source to identify the speaker of a passage written to be spoken.

These elements are formally defined as follows:

```
<!-- 6.3:  Paragraph-level chunks (cont'd)                    -->
<!-- (continuation of sec. 6)                                 -->
<!ELEMENT sp            - o  (spkr*, %pSeq)       -(sp)        >
<!ATTLIST sp                 %global;
          who                IDREF               #IMPLIED     >
<!ELEMENT spkr          - -  (#PCDATA)                         >
<!ATTLIST spkr               %global;                           >
```

6.4 Poems

Poems or fragments of verse or song may appear in either spoken or written texts. It is recommended, but not required, that they should be distinguished from other elements, using the <POEM> tag. This element may appear either between or within other chunks. It contains an optional series of <HEAD> elements followed by one or more <L> (for line) elements. The <L> element contains any sequence of phrase level elements; it is not however counted as a phrase-level element, since it can appear only within the <POEM> element, where its presence is required. Note that the <L> element is used to mark metrical lines, rather than typographic lines; if necessary, the <LB> tag can be used to mark the position of a line break in prose).

No provision is made for marking units of verse such as stanzas, verse paragraphs etc. An attribute PART is however available to mark part-lines in verse, where this is felt appropriate. In summary:

<POEM> a poem, or an extract from one, embedded or quoted within a spoken or written text.

<L> a line of verse. Attributes include:

PART indicates whether the verse line is metrically complete Values:

- *u* metricality is not marked or inapplicable
- *y* the line is metrically complete
- *n* the line is metrically incomplete

These elements are formally defined as follows:

```
<!-- 6.4:  Paragraph-level chunks (cont'd)                    -->
<!-- (continuation of sec. 6)                                 -->
<!ELEMENT poem          - o  (head*, epigraph?, l+)            >
<!ATTLIST poem               %global;                          >
<!ELEMENT l             - o  (%phSeq)                          >
<!ATTLIST l                  %global;
          part               (y|n|u)             u            >
```

6.5 Lists

Like poems, the tagging of lists is recommended but not required. Also like poems, the contents of lists are more tightly specified than those of other chunks, and lists may appear both within and between other chunks: for example, a textual division may consist only of lists, or the lists may appear within paragraphs. Unlike poems, lists are not permitted within spoken texts.

A list consists of an optional <HEAD> element, followed by one or more <ITEM> elements, each of which may optionally be prefixed by a <LABEL>element. In printed or written texts, lists are usually signalled by special layout conventions (such as those used in the present document). Of rather more importance is the fact that lists and their items generally exhibit quite different syntactic patterns from those of their surroundings.

The <LABEL> element is used to hold the identifier or tag sometimes attached to a list item, for example "(a)". It may also contain a word or phrase used for a similar purpose, as in the lists of tags given in the current document. It is a phrase-level element, and may thus appear outside lists, for example in situations where it is not possible or desirable to identify the list itself but it is desirable to distinguish labels from the rest of the text.

The <ITEM> element may appear only inside lists. It contains the same mixture of chunk elements as a paragraph, or a series of paragraphs. It may thus contain one or more nested lists. To avoid ambiguity it is therefore essential that end-tags be supplied for items and lists, as for other chunk elements.

In summary,

<LIST> a collection of distinct items flagged as such by special layout in written texts, often functioning as a single syntactic unit.

<ITEM> an item within a list.

<LABEL> an enumerator or other label attached to a list item or appearing freely within a text.

These elements are formally defined as follows:

```
<!-- 6.5:  Paragraph-level chunks (cont'd)                    -->
<!-- (continuation of sec. 6)                                 -->
<!ELEMENT list          - -  (head*, (label?, item)+)          >
<!ATTLIST list               %global;                          >
<!ELEMENT item          - -  (%seq)                            >
<!ATTLIST item               %global;                          >
<!ELEMENT label         - -  (%phSeq)                          >
<!ATTLIST label              %global;                          >
```

6.6 Citations

Bibliographic citations or references are very frequent in some kinds of writing, and almost entirely absent from others. It is useful to distinguish them from surrounding text, both because they have a very different linguistic structure, and for information retrieval purposes. The optional <CITN> element should contain the whole of any bibliographic citation distinguished in a text, including, for example, page and volume numbers.

Some constituents of the citation may also be tagged in their own right, using the appropriate phrase-level tags (<TITLE>, <PROPNAME>, <DATE> etc.) but CDIF does not require or enforce this. The <CITN> element contains any sequence of phrase-level elements. It is defined as follows:

```
<!-- 6.6:  Paragraph-level chunks (cont'd)                    -->
<!-- (continuation of sec. 6)                                 -->
<!ELEMENT citn          - -  (%seq)                             >
<!ATTLIST citn               %global;                           >
```

6.7 Notes

The <NOTE> element is used both for original foot, end or side notes, where these have been included in a transcribed text; and for any comment or additional explanation supplied by the transcriber, for example relating to some part of the text for which it is not clear which CDIF tag is the most appropriate. The two usages are distinguished by the TYPE attribute.

Original notes may contain any mixture of other chunks, and may also contain paragraphs: they may appear in written texts only. Transcribers' notes may contain #PCDATA only and can appear in either written or spoken text.

<NOTE> a foot-, end- or side-note in a written text, not forming part of the main text; any form of additional comment or gloss in either written or spoken texts. Attributes include:

TYPE identifies the provenance of the note i.e. editorial or authorial. Values:

ed note is supplied by transcriber or proof-reader

orig note is in the original

ED code for the person or organisation responsible for an editorial note.

PLACE for a written text, specifies the location of an original note in the source text. Values:

foot note at foot of page in original.

end note at end of current division or text in original.

side note in left or right margin of original

unspec original placement of note unknown or unspecified.

The formal definition for this element is as follows:

```
<!-- 6.7:  Paragraph-level chunks (cont'd)                    -->
<!-- (continuation of sec. 6)                                 -->
<!ELEMENT note          - -  (%seq)                             >
<!ATTLIST note               %global;
          ed                 CDATA               #IMPLIED
          place              (side|foot|end|unspec)
                                                 #IMPLIED
          type               (ed|orig)           #IMPLIED      >
```

7 Phrase level elements

Phrase level elements may appear anywhere that character data is permitted. For written texts, this implies that they cannot normally appear directly within a division, but must be contained with some other element. For spoken texts, this is not the case.

In the following discussion, the various phrase level elements have been grouped by function as follows:

- editorial changes

- highlighted phrases
- features of spoken texts
- miscellaneous

7.1 Addition, deletion and regularisation

Four optional tags are provided to record editorial changes made to the texts being transcribed. Transcribers may elect to make such changes silently, to use these tags only for changes about which there is a degree of doubt or to record all such changes. The header should indicate which policy has been adopted, where possible.

These tags may appear in both spoken and written texts. The following kinds of editorial change may be distinguished:

addition for example, of material which is present in the original but has been accidentally omitted during transcription

deletion for example, of material regarded as sensitive, irrelevant or untranscribeable

regularisation for example, of material believed to be erroneous or nonstandard in the original

non-regularisation for example, of material apparently erroneous or nonstandard in the original

Tags are provided for each of these kinds of change, as listed below. In each case the content of the element represents the result of the editorial change, i.e. the thing added for <ADD>, nothing at all for <DEL>, the regularized form for <REG> or the non-regularised form for <SIC>, and may contain any sequence of phrase level elements. Other aspects of the editorial change are recorded as attributes, as further detailed below:

<ADD> an editorial addition, supplying for example a word missed out unintentionally during transcription of a spoken or written text. Attributes include:

- ED identifies the person or organization responsible for the editorial decision.
- CAUSE describes the cause for the editorial change.

<DEL> an editorial deletion; marks the spot where the original source text has been deleted. Attributes include:

- ED identifies the person responsible for the editorial decision.
- CAUSE describes the cause for the editorial change.
- DESC brief description of the material deleted.

<REG> any editorial regularisation, e.g. to correct something mistranscribed or mis-spelled, or to normalise variant spellings. Attributes include:

- ED identifies the person or organization responsible for the editorial decision.
- CAUSE describes the cause for the editorial change
- SIC supplies the original form of whatever has been regularised

<SIC> a word or phrase which has not been regularised, but which is in doubt; for example, a spoken word which the transcribers cannot recognise, or a dubious spelling. Attributes include:

ED identifies the person or organization responsible for the editorial decision.

CAUSE describes the cause for the editorial change.

REG supplies a regularised form of the word or phrase.

These tags are formally defined as follows:

```
<!-- 7.1:  Phrase level elements                                  -->
<!ELEMENT add           - -  (%phSeq)                              >
<!ATTLIST add                %global;
          ed                 NAME                #IMPLIED
          cause              CDATA               #IMPLIED        >
<!ELEMENT del           - o  EMPTY                                 >
<!ATTLIST del                %global;
          ed                 NAME                #IMPLIED
          desc               CDATA               #IMPLIED
          cause              CDATA               #IMPLIED        >
<!ELEMENT reg           - -  (%phSeq)                              >
<!ATTLIST reg                %global;
          ed                 NAME                #IMPLIED
          cause              CDATA               #IMPLIED
          sic                CDATA               #IMPLIED        >
<!ELEMENT sic           - -  (%phSeq)                              >
<!ATTLIST sic                %global;
          ed                 NAME                #IMPLIED
          reg                CDATA               #IMPLIED
          cause              CDATA               #IMPLIED        >
```

7.2 Highlighted phrases

Typographic highlighting (use of quotation marks, italics, bold etc.) in written texts serves a wide variety of purposes, not all of them either self-evident or consistent. Where the boundaries of the highlighted matter coincide with those of some other CDIF element, no further tags should be introduced. Instead, for quoted matter, the quotation marks should be preserved, and for highlighted matter, the rendition attribute should be supplied. Where this is not the case, the <HI> element may be used where preserving the typographic distinction is felt to be necessary.

Quotation marks as such should always be represented by entity references within the text. The reference name used will depend on whether or not the usage of quotation marks in the text has been normalised. Information in the header should describe the course taken for a particular text. For candidate entity names, see section **??**. Where the quoted text is a true quotation (that is, a phrase or sequence attributed to someone other than the current narrator or writer) the <QUOTE> element discussed in section **??** may optionally be used. This does not apply to dialogue in fictional works, which is not marked, except by the presence of the quotation mark entities, in CDIF.

Highlighted phrases, and the kind of highlighting used, may be recorded in one of two ways:

- using the global R attribute
- using the phrase-level <HI> element

The former is appropriate where the function of the highlighting is clear, for example, to mark a heading, and where the boundaries of the highlighted phrase therefore coincide with the boundaries of some other CDIF element. The latter is appropriate when the function is not clear, where CDIF does not provide a tag to identify the feature concerned or where the highlighted phrase is not co-terminous with some other CDIF element.

Where the <HI> element is used, its R attribute must be supplied. On all other CDIF elements, the R attribute is optional. Legal values for it should be taken from the following short list, which may be extended as required

00pt pointsize ('00' should be two digits)

00ld leading ('00' should be two digits)

000m length of measure ('000' should be 3 digits)

bo bold face

bx boxed

it italic font

ql left aligned

qr right aligned

qc centred

qt quoted

ro roman typeface

sc small caps

st struck out

ul underlined

All numerical values above are assumed to be specified in points (i.e. 0.0138 inches). More than one value from this list may be supplied, in which case they should be joined by spaces, for example

```
<hi rend="bo it 12pt 16ld 480m">
```

for an element which is set to a 480 point measure, using 12 point type, in a bold italic font.

It should be noted that the purpose of the R attribute is *not* to provide information adequate to the needs of a typesetter, but simply to record such qualitative information about the original as is likely to be helpful in linguistic analysis.

Highlighted phrases must be entirely contained within some other CDIF element, like other phrase-elements. This implies that where, for example, a bolded passage contains more than one paragraph, or an italicised phrase begins in one verse line and ends in another, the <HI>element must be closed at the end of the enclosing element, and then re-opened within the next. For example, an italicised passage which crosses a paragraph boundary should be tagged as follows:

```
<p>This is the start of a paragraph which
  <hi r=it>switches to italics here and then
  goes on for several paragraphs</hi>
<p r=it>
   This second paragraph is all in italics
   and so should have no hi tag
<p><hi r=it>This is the last bit of italics</hi> and
    the rest is in roman.
```

In summary,

<HI> a passage of written text which is typographically highlighted for example by italics or bold, where the reason for this cannot be expressed by other tags.

The formal definition for the <HI> element is as follows:

```
<!-- 7.2:  Phrase level elements (cont'd)                    -->
<!-- (continuation of sec. 7.1)                              -->
<!ELEMENT hi            - -  (%phSeq)                         >
<!ATTLIST hi                 %global;                         >
```

7.3 Spoken phrase-level elements

A spoken text consists of an optional alignment map (see **??**) and *spSeq* , that is, a sequence of utterances and other phrase level items, either directly contained by the text or grouped into <DIV>s. The divisions of a spoken text (if any) contain any sequence of spoken phrase elements: there is no class of spoken element analoguous to the "chunks" of a written text.

7.3.1 Utterances

An utterance is a discrete sequence of speech produced by one participant in a conversation and is represented in CDIF as a <U> element. It may overlap with other utterances, or other events in the spoken text, and may contain any of the other spoken or common phrase-level elements, except <U>. It has a mandatory WHO attribute which identifies the person or group of people making the utterance, using a unique code defined in a section of the header.

In summary,

<U> an utterance by a single speaker. Attributes include:

WHO identifies the person or group responsible for the utterance.

The formal definition for the <U> element is as follows:

```
<!-- 7.3.1:  High level structure (spoken) (cont'd)          -->
<!-- (continuation of sec. 2.2)                              -->
<!ELEMENT u             - o  ( %spSeq )                       >
<!ATTLIST u                  %global;
          who                NAME                #REQUIRED    >
```

Rules for the transcription and normalisation of speech are further discussed in TGCW21 *Spoken Corpus Transcription Guide*. The editorial tags discussed in section **??** above may be used to represent normalisation practice when dealing with transcribed speech.

CDIF distinguishes the following kinds of para-linguistic phenomenon:

voice quality for example, tempo, pitch etc.: the points in which changes in these occur for a given speaker are marked with the <SHIFT> tag

non-verbal but vocalised sounds for example, coughs, humming noises etc. These are described when relevant, using the <VOCAL> tag

non-verbal and non-vocal events for example, passing lorries or gestures and actions by the participants. These are described, when relevant, using the <EVENT> tag

significant pauses both between and within utterances; these are indicated using the <PAUSE> tag.

To the above list should be added the <UNCLEAR> element, used to mark where the sound being transcribed cannot be interpreted and the <PTR> element used to indicate temporal alignment of other elements within a spoken text. Other aspects of spoken texts are not explicitly recorded in the encoding.

As noted in the following summary list, the DUR attribute may be used with several of these elements to indicate the duration of the phenomenon concerned in seconds and the DESC attribute to supply a brief verbal description of it.

<VOCAL> a non-linguistic but communicative sound made by one of the participants in a spoken text. Attributes include:

DUR duration of the sound in seconds.

DESC describes the kind of sound made.

<PAUSE> a marked pause during or between utterances in a spoken text. Attributes include:

DUR length of the pause in seconds.

<SHIFT> a marked change in voice quality for any one speaker. Attributes include:

NEW describes the voice quality after the shift.

<EVENT> a non-communicative event (e.g. a door slamming) occurring during a conversation and regarded as worthy of note. Attributes include:

DUR duration of the event in seconds.

DESC description of the event.

<UNCLEAR> a point in a spoken text at which it is unclear what is happening, e.g. who is speaking or what is being said. Attributes include:

DUR specifies the length of the passage in seconds.

<TRUNC> a word or phrase which has been truncated during speech. Attributes include:

ED identifies the person or organization responsible for the editorial decision.

REG supplies a regularised form of the word or phrase.

CAUSE describes the cause for the truncation.

These elements have the following formal declaration:

```
<!-- 7.3.1:  High level structure (spoken) (cont'd)          -->
<!-- (continuation of sec. 2.2)                              -->
<!ELEMENT vocal         - o  EMPTY                             >
<!ATTLIST vocal              %global;
          desc               CDATA               #IMPLIED
          dur                NUMBER              #IMPLIED      >
<!ELEMENT pause         - o  EMPTY                             >
<!ATTLIST pause              %global;
          dur                NUMBER              #IMPLIED      >
<!ELEMENT shift         - o  EMPTY                             >
<!ATTLIST shift              %global;
          new                CDATA               #IMPLIED      >
<!ELEMENT event         - o  EMPTY                             >
<!ATTLIST event              %global;
          desc               CDATA               #IMPLIED
          dur                NUMBER              #IMPLIED      >
<!ELEMENT unclear       - o  EMPTY                             >
<!ATTLIST unclear            %global;
          dur                NUMBER              #IMPLIED      >
<!ELEMENT trunc         - -  (#PCDATA)                         >
<!ATTLIST trunc              %global;
          ed                 NAME                #IMPLIED
          reg                CDATA               #IMPLIED
          cause              CDATA               #IMPLIED      >
```

7.3.2 Alignment of overlapping speech

By default it is assumed that the events represented in a transcription are non-overlapping and that they are transcribed in temporal sequence. That is, unless otherwise specified, it is implied that the end of one utterance precedes the start of the next following it in the text, perhaps with an interposed <PAUSE> element. Where this is not the case, the following mechanism is used.

For each point of synchrony, i.e. at each place where the number of simultaneous utterances, events, vocals etc. increases or decreases, a <LOC> element is defined within an <ALIGN> element, which appears at the start of the enclosing <DIV>, if any. At each place to be synchronised within the text, a <PTR>element is inserted. The T (target) attributes of these <PTR> elements are then used to specify the identifier of the <LOC> with which each is to be synchronised.

For example, consider the following dialogue, represented first as it might appear in a conventional playscript:

```
Tom: I used to smoke --
Bob: (interrupting) You used to smoke?
Tom: (at the same time) a lot more than this.
But I never inhaled the smoke
```

This would appear in the Longman's data capture format something like the following:

```
<1> I used to smoke [ a lot more than this ]
<2>                 [ you used to smoke]
<1> but I never inhaled the smoke
```

In CDIF this would be rendered as follows:

```
<align><loc ID=P1><loc ID=P2></align>
  ....
<u who=TOM>I used to smoke <ptr t=P1> a lot more than this
```

```
      <ptr t=P2>but I never inhaled the smoke
  <u who=BOB><ptr t=P1>You used to smoke<ptr t=P2>
```

Note that although Bob's utterance follows Tom's sequentially in the text, it is aligned temporally with its middle, without any need to disrupt the normal syntax of the text.

The formal declarations for the elements used to implement this alignment mechanism are as follows:

```
<!-- 7.3.2:  High level structure (spoken) (cont'd)          -->
<!-- (continuation of sec. 2.2)                              -->
<!ELEMENT align         - -  (loc+)                           >
<!ATTLIST align              %global;                         >
<!ELEMENT ptr           - o  EMPTY                            >
<!ATTLIST ptr                %global;
          t                  IDREF              #REQUIRED     >
<!ELEMENT loc           - o  EMPTY                            >
<!ATTLIST loc                %global;                         >
```

7.4 Miscellaneous phrase-level elements

In addition to those already discussed, there is a small number of miscellaneous phrase-level elements. Most of them appear in written texts only, at any point that phrase-level elements are allowed; <ABBREV>, <PROPNAME> and <TITLE> may also appear in spoken texts; <TRUNC> is restricted to spoken texts only. Brief descriptions of each of them follow:

<ABBREV> any acronym or abbreviation.

<DATE> a calendar date in any format.

<DISTINCT> a word or phrase which is markedly distinguishable from the surrounding text, for example because it is non-English, technical, archaic, regional etc. Attributes include:

TYPE type of distinction identified Values:

- *r* regional
- *t* technical
- *f* foreign
- *u* unspecified
- *a* archaic

<PB> marks the start of a new page in the original source; used to indicate where e.g. articles in periodicals are split across several pages.

<LB> marks the start of a new (printed) line in the original source.

<PROPNAME> proper name of a person, place or institution.

<SALUTE> a formulaic greeting appearing at the start or the end of a spoken or a written text.

<STAGE> contains any kind of stage direction within a dramatic text. Attributes include:

TYPE indicates the kind of stage direction. Values:

- *m* movement, i.e. an exit or entrance
- *x* mixed, i.e. more than one of the above

d delivery, e.g. shouting

s setting, i.e. describing the setting or scene

u unknown or unspecified

a action, i.e. a piece of business

<TITLE> the title of a book, song or similar bibliographic entity appearing anywhere within a text.

Formal declarations for these elements follow:

```
<!-- 7.4:  Phrase level elements (cont'd)                    -->
<!-- (continuation of sec. 7.1)                              -->
<!ELEMENT abbrev        - -  (%phSeq)                         >
<!ATTLIST abbrev             %global;                         >
<!ELEMENT date          - -  (%phSeq)                         >
<!ATTLIST date               %global;                         >
<!ELEMENT distinct      - -  (%phSeq)                         >
<!ATTLIST distinct           %global;
          type               (f|a|t|r|u)         u            >
<!ELEMENT pb            - o  EMPTY                            >
<!ATTLIST pb                 %global;                         >
<!ELEMENT lb            - o  EMPTY                            >
<!ATTLIST lb                 %global;                         >
<!ELEMENT propName      - -  (%phSeq)                         >
<!ATTLIST propName           %global;                         >
<!ELEMENT salute        - -  (%phSeq)                         >
<!ATTLIST salute             %global;                         >
<!ELEMENT stage         - -  (%phSeq)                         >
<!ATTLIST stage              %global;
          type               (m|s|a|d|x|u)       u            >
<!ELEMENT title         - -  (%phSeq)                         >
<!ATTLIST title              %global;                         >
```

7.5 Segmentation

Texts which have been processed by the CLAWS system will be enriched in two ways. Firstly, they will be segmented into syntactic units resembling sentences [7]. Secondly, each token will be associated with a "word class" code, as discussed in section **??** below.

Segmentation is represented using the general purpose <S> element. Each of the chunks defined in section **??** above, will be decomposed into a sequence of <S> elements, within which other phrase-level tags will be entirely contained. Once this segmentation has been performed, the <S> element will become the basic organizational principle for the whole corpus. For texts included in the "core" corpus of analysed texts, the <S> element will additionally carry an automatically-assigned syntactic code, and may self-nest.

Because the segmentation is carried out as a separate stage in the processing of the texts, it will be necessary to maintain two versions of the CDIF dtd. In the first version, used to validate texts before segmentation is carried out, the *sequence elements* discussed in section **??** have their default definitions, for example paragraphs may contain any phrase-level elements. In the second version, used to validate texts after segmentation is carried out, the same elements are redefined to permit only sequences of <S> elements. To achieve this, the parameter elements *seq* , *phSeq* and *spSeq* are redefined as follows:

[7] The segmentation principles embodied in the CLAWS program are defined in Garside (1987); see bibliography

```
<!-- 7.5:                                                    -->
<!-- Redefine content model for common sequences             -->
<!ENTITY % seq '(s+)' >
<!ENTITY % phSeq '(s+)' >
<!ENTITY % spSeq '(s+)' >
```

These redefinitions should be embedded within the doctype subset associated with a file to be parsed, as in the following example:

```
<!DOCTYPE cdif SYSTEM "cdif.dtd" [
<!ENTITY % seq "(s+)">
<!-- etc -->
<!ENTITY f1 SYSTEM "data.1">
<!ENTITY f2 SYSTEM "data.2">
<!-- declarations for other data files -->
]>
<cdif>
&f1;
&f2;
</cdif>
```

Once segmented, texts are allocated an identifier, as discussed in TGCW34. A number, unique within each text, is used initially for this and forms the value of the N attribute on each <S> element. At a later stage, this number will be prefixed with a code for the text to provide an identifier unique within the whole corpus, which can be used as a value for the global IDattribute on every <S> element.

A value must be supplied for the TYPE attribute on all <S> elements which appear in texts which have been included in the *core* linguistically-analysed corpus, i.e. those texts on which Lancaster have performed a *skeletal parsing*. This value identifies the results of the skeletal parse, and is a code taken from the following list:

(list to be supplied)

The <S> element has the following formal definition:

```
<!-- 7.5:  Phrase level elements (cont'd)                   -->
<!-- (continuation of sec. 7.1)                             -->
<!ELEMENT s             - o  (%phSeq)                        >
<!ATTLIST s                  %global;
          type               CDATA              #IMPLIED     >
```

8 Entities

General entity references are used for two distinct purposes in the CDIF scheme. Firstly, they are used to represent the word-class codes assigned by the CLAWS program, as further discussed in the next section and in various other working papers. Secondly, they are used, as is normal SGML practice, to represent characters or symbols not available from the default character reportoire.

8.1 Wordclass entities

Working paper TGDW08 gives full details and definitions for the CLAWS C5 wordclass codes to be added to CDIF texts, but a summary list is provided here for convenience.

Each wordclass code will be represented by an entity reference using the same name, which will be suffixed to the relevant word within the text, ahead of any punctuation. For example, the sentence "The cat sat on the mattress." might appear in CDIF as

```
<s>The&AT0; cat&NN1; sat&VVD; on&PRP;
   the&AT0; mattress&NN1;.&PUN;
```

[8]

A set of entity declarations allowing each such reference to be replaced by a null string will be used during testing. At a later stage, alternative sets of declarations may be provided, in which each wordclass entity reference will be replaced by a <PTR> element, the target of which will be a TEI-conformant *feature structure* for the word class definition itself. A draft set of such feature structures is available in document TGDW09.

The following wordclass codes will be used for the majority of texts in the corpus :

AJ0 adjective (unmarked) e.g. *good, old*

AJC comparative adjective e.g. *better, older*

AJS superlative adjective e.g. *best, oldest*

AT0 article e.g. *the, a, an*

AV0 adverb (unmarked) e.g. *often, well, longer, furthest*

AVP adverb particle e.g. *up, off, out*

AVQ wh-adverb e.g. *when, how, why*

CJC coordinating conjunction e.g. *and, or*

CJS subordinating conjunction e.g. *although, when*

CJT the conjunction *that*

CRD cardinal numeral e.g. *3, fifty-five, 6609* (excluding *one*)

DPS possessive determiner form e.g. *your, their*

DT0 general determiner e.g. *these, some*

DTQ wh-determiner e.g. *whose, which*

EX0 existential *there*

ITJ interjection or other isolate e.g. *oh, yes, mhm*

NN0 noun (neutral for number) e.g. *aircraft, data*

NN1 singular noun e.g. *pencil, goose*

NN2 plural noun e.g. *pencils, geese*

NP0 proper noun e.g. *london, michael, mars*

ONE the word *one* (including numeral and non-numeral uses)

ORD ordinal e.g. *sixth, 77th, last*

PNI indefinite pronoun e.g. *none, everything*

[8] SGML does not in fact require any semicolon other than the one following the second "NN1" (because no white space separates this token from the following punctuation); they are retained here for convenience of processing by the CLAWS system.

PNP personal pronoun e.g. *you, them, ours*

PNQ wh-pronoun e.g. *who, whoever*

PNX reflexive pronoun e.g. *itself, ourselves*

POS the possessive (genitive) morpheme *'s* or '

PRF the preposition *of*

PRP preposition (except for *of*) e.g. *for, above, to*

TO0 infinitive marker i.e. *to*

UNC "unclassified" items which are not words of the English lexicon or do not belong to any recognized category. e.g. formulae, such as "XX61"; foreign words; *both* when correlative with *and*; etc.

VBB the base forms of the verb "be", except infinitive, i.e. *am, are*

VBD past form of the verb "be", i.e. *was, were*

VBG -ing form of the verb "be", i.e. *being*

VBI infinitive of the verb "be"

VBN past participle of the verb "be", i.e. *been*

VBZ -s form of the verb "be", i.e. *is, 's*

VDB base form of the verb "do", except the infinitive

VDD past form of the verb "do", i.e. *did*

VDG -ing form of the verb "do", i.e. *doing*

VDI infinitive of the verb "do"

VDN past participle of the verb "do", i.e. *done*

VDZ -s form of the verb "do", i.e. *does*

VHB base form of the verb "have", except the infinitive

VHD past tense form of the verb "have", i.e. *had, 'd*

VHG -ing form of the verb "have", i.e. *having*

VHI infinitive of the verb "have"

VHN past participle of the verb "have", i.e. *had*

VHZ -s form of the verb "have", i.e. *has, 's*

VM0 modal auxiliary verb e.g. *can, could, will, 'll*

VVB base form of lexical verb, except the infinitive e.g. *take, live*

VVD past tense form of lexical verb e.g. *took, lived*

VVG -ing form of lexical verb e.g. *taking, living*

VVI infinitive of lexical verb

VVN past participle form of lexical verb e.g. *taken, lived*

VVZ -s form of lexical verb e.g. *takes, lives*

XX0 the negative *not* or *n't*

ZZ0 alphabetical symbol e.g. *A, b, c, d*

AJ0-AV0 adjective or adverb

AV0-AJ0 adverb or adjective

AJ0-NN1 adjective or singular common noun

NN1-AJ0 singular common noun or adjective

AJ0-VVD adjective or past tense verb

VVD-AJ0 past tense verb or adjective

AJ0-VVG adjective or -ing from of the verb

VVG-AJ0 -ing from of the verb or adjective

AJ0-VVN adjective or past participle

VVN-AJ0 past participle or adjective

AVP-PRP adverb particle or preposition

PRP-AVP preposition or adverb particle

CJS-PRP subordinating conjunction or preposition

PRP-CJS preposition or subordinating conjunction

NN1-NP0 singular common noun or proper noun

NP0-NN1 proper noun or singular common noun

NN1-VVG singular common noun or -ing form of the verb

VVG-NN1 -ing form of the verb or singular common noun

VVD-VVN past tense verb or past participle

VVN-VVD past participle or past tense verb

PUL left bracket (i.e. (or [ )

PUN punctuation mark - normal (i.e. . ! , : ; - ? ...)

PUQ quotation mark (i.e. '' " ")

PUR right bracket (i.e.) or])

For texts which have been included in the *core* corpus, i.e. those with enhanced linguistic analysis, a different and much enlarged set of word class codes will be used, known as the CLAWS 2B tagset: a summary list of these codes is given below:

APP possessive pronoun, pre-nominal (my, your, our)

AT article (the, no)

AT1 singular article (a, an, every)

BCS before-conjunction (in order, even, preceding that, if etc)

BTO before-infinitive marker (in order, so as, preceding to)

CC coordinating conjunction (and, or)

CCB coordinating conjunction (but)

CS subordinating conjunction (if, because, unless)

CSA as as conjunction

CSN than as conjunction

CST that as conjunction

CSW whether as conjunction

DA after-determiner (capable of pronominal function) (such, former, same)

DA1 singular after-determiner (little, much)

DA2 plural after-determiner (few, several, many)

DA2R comparative plural after-determiner (fewer)

DA2T superlative plural after-determiner (fewest)

DAR comparative after-determiner (more, less)

DAT superlative after-determiner (most, least)

DB before determiner (capable of pronominal function) (all, half)

DB2 plural before-determiner (capable of pronominal function) (both)

DD determiner (capable of pronominal function) (any, some)

DD1 singular determiner (this, that, another)

DD2 plural determiner (these,those)

DDQ wh-determiner (which, what)

DDQ$ wh-determiner, genitive (whose)

DDQV wh-ever determiner, (whichever, whatever)

EX existential

IF for as preposition

II preposition

IO of as preposition

IW with, without as prepositions

JJ general adjective

JJR general comparative adjective (older, better, stronger)

JJT general superlative adjective (oldest, best, strongest)

JK adjective catenative (able in be able to, willing in be willing to)

LE leading coordinator (both in both..and, either in either..or)

MC cardinal number,neutral for number (two, three..)

MC$ genitive cardinal number, neutral for number (10's, 100's)

MC-MC hyphenated number (40-50, 1770-1827)

MC1 singular cardinal number (one)

MC2 plural cardinal number (tens, hundreds)

MD ordinal number (first, second, next, last)

MF fraction,neutral for number (quarters, two-thirds)

ND1 singular noun of direction (north, southeast)

NN common noun,neutral for number (sheep, cod, headquarters)

NN1 singular common noun (book, girl)

NN2 plural common noun (books, girls)

NNJ organization noun, neutral for number (co., group)

NNJ2 organization noun, plural (groups, councils, unions)

NNL locative noun, neutral for number (is.)

NNL1 singular locative noun (island, street)

NNL2 plural locative noun (islands, streets)

NNO numeral noun, neutral for number (dozen, hundred)

NNO2 numeral noun, plural (hundreds, thousands)

NNSA1 following noun of style or title, abbreviatory (m.a.)

NNSA2 following plural noun of style or title, abbreviatory

NNSB preceding noun of style or title, abbreviatory (rt.hon.)

NNSB1 prec. sing. noun of style or title, abbreviatory (prof.)

NNSB2 prec. plural noun of style or title, abbreviatory (messrs.)

NNT1 temporal noun,singular (day, week, year)

NNT2 temporal noun,plural (days, weeks, years)

NNU unit of measurement,neutral for number (in, cc)

NNU1 singular unit of measurement (inch, centimetre)

NNU2 plural unit of measurement (ins., feet)

NP proper noun, neutral for number (indies, andes)

NP1 singular proper noun (london, jane, frederick)

NP2 plural proper noun (londons, johns, marys)

NPD1 singular weekday noun (sunday)

NPD2 plural weekday noun (sundays)

NPM1 singular month noun (october)

NPM2 plural month noun (octobers)

PN indefinite pronoun, neutral for number (none)

PN1 indefinite pronoun, singular (anyone, everything, nobody, one)

PNQO whom

PNQS who

PNQV$ whosoever

PNQVO whomever

PNQVS whoever

PNX1 reflexive indefinite pronoun (oneself)

PP$ nominal possessive personal pronoun (mine, yours)

PPH1 it

PPHO1 him, her

PPHO2 them

PPHS1 he, she

PPHS2 they

PPIO1 me

PPIO2 us

PPIS1 I

PPIS2 we

PPX1 singular reflexive personal pronoun (yourself, itself)

PPX2 plural reflexive personal pronoun (yourselves, themselves)

PPY you

RA adverb, after nominal head (else, galore)

REX adverb introducing appositional constructions (namely, e.g.)

RG degree adverb (very, so, too)

RGA post-adjectival/adverbial degree adverb (enough, indeed)

RGQ wh- degree adverb (how)

RGQV wh-ever degree adverb (however)

RGR comparative degree adverb (more, less)

RGT superlative degree adverb (most, least)

RL locative adverb (alongside forward)

RP prep. adverb, also particle (about, in)

RPK prep. adv., catenative (about in be about to)

RR general adverb

RRQ wh- general adverb (where, when, why, how)

RRQV wh-ever general adverb (wherever, whenever)

RRR comparative general adverb (better, longer)

RRT superlative general adverb (best, longest)

RT nominal adverb of time (now, tomorrow)

TO infinitive marker (to)

UH interjection (oh, yes, um)

VBDR were

VBDZ was

VBG being

VBI be infinitive. To be or not... It will be ...

VBM am

VBN been

VBR are

VBZ is

VD0 do - finite non-3rd pers sing

VDD did

VDG doing

VDI infinitive do: I may do; To do...

VDN done

VDZ does

VH0 have - finite non-3rd pers sing

VHD had (past tense)

VHG having

VHI have infin.

VHN had (past participle)

VHZ has

VM modal auxiliary (can, will, would)

VMK modal catenative (ought, used)

VV0 Non-3PS form of lexical verb (give, work)

VVD past tense of lexical verb (gave, worked)

VVG -ing participle of lexical verb (giving, working)

VVGK -ing participle catenative (going in be going to)

VVI infinitive - (to) SEE. It will APPEAR

VVN past participle of lexical verb (given, worked)

VVNK past participle catenative (bound in be bound to)

VVZ s- form of lexical verb (gives, works)

XX not

ZZ1 singular letter of the alphabet

ZZ2 plural letter of the alphabet (as, bs)

8.2 Character code entities

A full discussion of the entity names to be used for representing accented characters and other special symbols within all CDIF texts is provided in working paper TGCW25.

9 The header

A discussion of the intended contents of the header and of its relation to the recently-published TEI proposals form the subject of TGCW34. The following dummy declaration is used for transition purposes.

```
<!-- 9:  The header                                          -->
<!ELEMENT header        - o  (#PCDATA)                        >
<!ATTLIST header             %global;                         >
```

10 Alphabetical list of elements in the CDIF scheme

To be supplied

11 The CDIF Document Type Declaration

```
\include{cdif.dtd}
```