TGCW34
The Relationship Between the TEI.2 Header and the BNC Corpus and Text Headers

Dominic Dunlop

Draft of 4th September, 1992

Contents

1 Introduction

This document sets out all the tags defined by [13, II:22, The TEI Header], and states if and how they are used in the BNC text and corpus headers. Where necessary, tags defined elsewhere in [13] are also referenced. Tags which may occur at multiple points in the header are referenced as many times as necessary.

1.1 Notation

The subsections which follow have the following layout:

n `<TEI.2-tagname>` `<CDIF-tagname>`

Corpus header Description of tag usage (if any) in corpus header

Text header Description of tag usage (if any) in text header

Notes The number *n* gives the "nesting level" of the tag: thus, the outermost tag (`TeiHeader`) is at level 1, and contains tags at level 2. These in turn may contain tags at level 3, and so on. The order in which tags are presented is generally "depth first"; that is, the order in which they would be encountered if actually reading through a CDIF corpus or text header. This contrasts with the generally "breadth first" approach of [13], which tends to present all the tags at a given level before moving on to any subordinate tags.

The information at the left of the page gives the TEI tagname, while the corresponding proposed CDIF tagname appears at the right of the page. Where the CDIF tag name differs from that of the TEI tag, the reason is to keep the length of the name down to eight or fewer characters, so keeping within the limitations of some SGML-processing software. The alternative name is taken from [16] if possible. CDIF tagnames corresponding to all TEI tagnames are given, even where this document recommends that a particular TEI tag is not used in CDIF. In descriptions of tag usage, the TEI tagname is used, even where a different tagname is proposed for BNC use.

In descriptions of tag usage and content, tag and attribute names, and possible attribute values, always appear in `typewriter font`. Where the precise content of an element is given — for example, `British National Corpus` — this also appears in typewriter font. *Italic font* is used to indicate a substitutable value, as in *title of source text*.

2 The TEI Header

1 `<TeiHeader>` `<header>`

Corpus header Contains CDIF corpus header information and occurs once at start of the corpus. The `type` attribute has the value `corpus`. The `creator` attribute identifies the person or institution responsible for the creation of the corpus header. The `status` attribute has the value `new` for release 1.0 of the BNC; `update` for subsequent releases. (See also `<editionStmt>`, §3.2.) The `datCreat` attribute (corresponding to TEI `dateCreated`) gives the date on which the header was created. The `datUpdat` attribute (TEI `dateUpdate`) is not initially used, but can be used subsequently to give the date on which an updated version of the header was created.

Text header Contains CDIF text header information and occurs at the beginning of each text. The `type` attribute has the value `text`. Other attributes are used as described for the corpus header.

Notes See `<fileDesc>` (§3), `<encodingDesc>` (§4), `<profileDesc>` (§5), and `<revisionDesc>` (§6) for subordinate elements.

3 The File Description

2 `<fileDesc>` `<fileDesc>`

Corpus header Contains information about the BNC as a whole.

Text header Contains information about an individual text within the BNC.

Notes See `<titleStmt>` (§3.1), `<editionStmt>` (§3.2), `<extent>` (§3.3), `<publicationStmt>` (§3.4), `<seriesStmt>` (§3.5), `<notesStmt>` (§3.6) and `<sourceDesc>` (§3.7) for subordinate elements.

3.1 The Title Statement

3 `<titleStmt>` `<titlStmt>`

Corpus header Bibliographic description of the BNC as a whole

Text header Bibliographic description of an individual electronic text

4 `<title>` `<title>`

Corpus header `The British National Corpus`

Text header *title of source text*`: an electronic sample` (or similar). In the event that a written source text has no obvious title, the title for the electronic text will be derived from that created for the source text (see `<citn.struct>`, §3.8). Titles for spoken texts will be generated by BNC staff. These titles will be unique across the corpus. Square brackets (`[]`, or rather `[` and `]`), will enclose generated titles: for example, `[Official leaflets A]: an electronic version`.

4 `<author>` `<author>`

Corpus header Not used

Text header Not used. (Author(s) if the source text are listed in the `<sourceDesc>` — see `<citn.struct>`, §3.8.)

4 `<sponsor>` `<sponsor>`

Corpus header `The British National Corpus Consortium`

Text header `The British National Corpus Consortium`

4 `<funder>` `<funder>`

Corpus header Some form of words about the constitution of the consortium and its sources of funding — the DTI, SERC and the commercial partners.

Text header `See information in corpus header`

4 `<principal>` `<princpl>`

Corpus header Not used

Text header Not used

4 `<resp>` `<resp>`

Corpus header Occurs once for each consortium member organization involved with all texts, giving name and rôle. (That is, Lancaster and OUCS.)

Text header Occurs once, giving name and rôle, for each consortium member organization involved with the text, in addition to those listed in the corpus header. (Either Chambers, Longman, or OUP, depending on who did the data capture.)

5 `<role>` `<role>`

Corpus header See `<resp>`.

Text header See `<resp>`.

5 `<name>` `<name>`

Corpus header See `<resp>`.

Text header See `<resp>`.

3.2 The Edition Statement

3 `<editionStmt>` `<edinStmt>`

Corpus header Gives version information for corpus as a whole. The `n` attribute has value the *version.revision*, where *version* changes if texts are added to or removed from the corpus, and *revision* changes if amendments are made within texts or the corpus header. The tag has no content.

Text header Gives version information for an individual electronic text. The `n` attribute has the value *revision*. This changes when the individual text is amended. The tag initially has no content, but descriptive prose content may be added for revisions after the first.

4 `<edition>` `<edition>`

Corpus header Not used

Text header Not used

4 `<resp>` `<resp>`

Corpus header Not used

Text header Not used

Notes The intention is that, when first generally released, the BNC as a whole should be at revision 1.0, and each text in the corpus at version 1. (We may use a different numbering scheme during development.) Under the scheme described, the *revision* part of the corpus release number changes whenever a new release of the corpus contains new revisions of one or more texts. Clearly, if a new release is produced each time any text changes, the corpus revision number could increase rapidly. Consequently, it is suggested that changed texts be batched, and are incorporated into the corpus no more frequently than every six months or so.

The reference scheme (see §4.5) requires that full references to objects in the BNC include the corpus edition to which the reference applies. This suggests that the edition information for the corpus (and hence the contents of the corpus itself) should not be changed frequently or lightly. There is also an implication that it should possible to regenerate any version of the corpus at any time, in order that references to a non-current version of the corpus may be satisfied. This issue should be considered when distribution arrangements for the corpus are made.

See also `<TeiHeader>` (§2), which records date information relating to revisions, and `<revisionDesc>` (§6), which records details of the changes made between revisions.

3.3 The Extent Statement

3 `<extent>` `<extent>`

Corpus header At a minimum, gives number of words in the corpus, and a definition of the term "words". May get much more verbose, giving number of texts, types of texts, words per type and so on. (But see also `<textClass>`, §5.3.)

Text header Gives number of words in the text, and the number of kilobytes (multiples of 1,024 octets, rounded up to the next integer) in its canonical CDIF representation as a UNIX text file using the ISO 646 coded character set. (The latter is useful in calculating media requirements or file download times.)

3.4 The Publication Statement

3 `<publicationStmt>` `<publStmt>`

Corpus header Contains information about distribution of the corpus as a whole.

Text header Contains information about the distribution of an individual electronic text, about the agency or agencies from which permission has been obtained for the inclusion of the text in the corpus, and about constraints attaching to that permission.

4 `<publisher>` `<publ>`

Corpus header Not used

Text header Not used

4 `<distributor>` `<distribr>`

Corpus header `The British National Corpus Consortium`

Text header `The British National Corpus Consortium`

4 `<authority>` `<authty>`

Corpus header Not used

Text header Not used

4 `<place>` `<place>`

Corpus header Not used

Text header Not used

4 `<address>` `<address>`

Corpus header Full contact information for BNC — presumably an address c/o OUCS. (Components, which include `<phone>`, `<fax>`, and `<email>` tags, not shown.)

Text header Full contact information for BNC, as in corpus header. This information is repeated in each text in order that a person who gets sight of an individual text, by whatever means, can know where to apply in order to obtain the whole corpus in the correct manner.

4 `<idno>` `<idno>`

Corpus header BNC_*version.revision*. (See also `<editionStmt>`, §3.2.) The `type` attribute has the value `local`.

Text header *six-character mixed-case filename*. The `type` attribute has the value `local`.

4 `<availability>` `<availty>`

Corpus header Describes availability of corpus as a whole, distinguishing between academic and commercial use, and between UK, Europe and the world as appropriate. The `status` attribute has the value `restricted`.

Text header Contains two elements. The first, as a series of paragraphs (`<p>`s), is a standard form of words stating that text may be used only by registered users of the corpus, who are individually responsible for protecting the associated intellectual property rights. Although this element is repeated in each text header, it should *not* be replaced by an entity reference in order to save space: it is important that the restrictions on the use of corpus material appear as plain text in each text header.

The second element is a list or series of lists as specified below. (Actually lists inside a `<p>`, so as to conform to the model specified in [13].) There may be multiple lists if an individual text embodies separate analytic works for which permissions are controlled by different agencies.

The `status` attribute has the value `restricted`.

5 `<list>` `<list>`

Corpus header Not used

Text header Lists organizations or individuals from whom permission has been obtained for inclusion of the work in the corpus, and any limits and restrictions associated with those permissions.

6 `<head>` `<head>`

Corpus header Not used

Text header Where a text embodies separate analytic works with individual permissions, indicates the analytic work or works to which the list applies. Not used if identical permissions apply to a whole text. Where separate permissions apply to separate parts of a text, a permissions list which does not refer to a specific part of the text refers to all those parts not covered explicitly by other lists.

7 `<ptr>` `<ptr>`

Corpus header Not used

Text header t attribute references <title.piece> to which permissions in the list apply. May also reference a <title> or <title.series> element for a <broadcast> if it is necessary to describe both rights to a broadcast and to the script or scripts used in that broadcast. (See §3.10.)

May occur more than once if identical permissions apply to more than one work.

6 <label> <label>

Corpus header Not used

Text header Subordinate elements give the name and location of an individual permission grantor.

7 <name> <name>

Corpus header Not used

Text header Name of the grantor.

7 <place> <place>

Corpus header Not used

Text header City, town or village of grantor. (See note below.)

6 <item> <item>

Corpus header Not used

Text header Territory for which permission is granted (normally world). If the permission granted varies from the default permissions for the corpus as a whole (set out in the corpus header <permissions> element), the restrictions on use should be described here.

4 <date> <date>

Corpus header Date this copy of the BNC header was created (if technically feasible).

Text header Date this copy of this text was created (if technically feasible).

Notes [13] requires that one of <publisher>, <distributor> or <authority> appears in <publicationStmt>. Of these, <distributor> seems best to describe the rôle played by the BNC consortium, both for the corpus as a whole and for its individual texts.

[13] suggests that further dates (for example, of data capture) be given as <note>s if the information is available. I would suggest that, if we record this information at all, it goes in the revision description (see §6).

The contents of <permission> are entirely legal according to [13], and consequently have the appearance of a square (or structured) peg rammed into a round, unstructured, hole. A sample element for an individual text would look something like:

```
<permission>
   <p>
     Unless otherwise stated below, or in an individual
     agreement signed by an officer of the British
     National Corpus Consortium, this text may be used
     anywhere in the world only for bona fide research
     purposes by registered users of the British
```

```
      National Corpus, who are jointly and severally
      responsible...
    <p>
    <list>
      <label>
        <name>Omnibus Press</name>
        <place>London</place>
      </label>
      <item>World</item>
    </list>
    <list>
      <head><ptr t=tp003641><ptr t=tp003643></head>
      <label>
        <name>Contact Press</name>
        <place>Toronto</place>
      </label>
      <item>World</item>
    </list>
    <list>
      <head><ptr t=tp003642><ptr t=tp003644></head>
      <label>
        <name>Music Sales Ltd.</name>
        <place>London</place>
      </label>
        <item>Rest of world</item>
      <label>
        <name>Columbia Records Inc.</name>
        <place>New York</place>
      </label>
      <item>North and South America<item>
    </list>
    ...
  </permission>
```

(The example is from the text `LCohen`, *Leonard Cohen, Prophet of the Heart*. In practice, we would not be likely to go into such detail.)

Note that we do not give `<address>`es for permissions grantors, even though we may hold those addresses on file.

3.5 The Series Statement

3 `<seriesStmt>` `<serStmt>`

Corpus header Not used. (The BNC is not a serial publication).

Text header Not used. (Where the source of an individual text is an issue of a serial publication, this is recorded in the source description — see §3.7).

3.6 The Notes Statement as an Element of the File Description

3 `<notesStmt>` `<noteStmt>`

Corpus header Not used

Text header Not used

4 <note> <note>

Corpus header Not used

Text header Not used

Notes While I can't for the moment see any use for <note>s in this context, I'm sure we'll need them, so the CDIF DTD should allow them.

3.7 The Source Description

2 <sourceDesc> <srcDesc>

Corpus header Not used

Text header Contains information about the source from which a text was derived.

Notes See <citn> (§3.8), <scriptStmt> (§3.9) and <recordingStmt> (§3.10) for subordinate elements.

3.8 The Citation

3 <citn> <citn>

Corpus header Not used

Text header Not used

3 <citn.struct> <cit.str>

Corpus header Not used

Text header Not used here for spoken texts (see <scriptStmt>, §3.9 and <recordingStmt>, §3.10). For written texts, gives as complete a bibliographic description as is practicable. Descriptions of any distinct analytic texts (newspaper stories, ephemeral items, learned papers etc.) are included by means of <title.piece> elements. A description of any series of which the text is a part is included by means of <title.series> elements. (See following elements.)

4 <title.piece> <title.p>

Corpus header Not used

Text header Gives the title of an individual analytic text which is part of a monographic source text. Repeated for each distinct analytic text. This and the following elements (<author>, <editor>, <imprint>, <idno> and <note>) are used only if the source text comprises one or more distinct parts for which the contents of one or more of the following elements varies. Thus, if, for example, a book consists of a number of short stories or papers by the same author, it is unnecessary to describe each with a <title.piece> unless to use a subsequent element to detail some variable aspect of the texts, such as their initial dates of publication. Examples of source texts for which <title.piece> might be used are newspapers or magazines containing articles; conference proceedings or journals containing papers; books containing short stories; and collections of ephemeral items such as leaflets.

Where an analytic text is distinct, but has no obvious title — as might be the case for a business letter or a hand-bill — a title is generated by BNC staff. Square brackets ([]) indicate a generated title, as in [Christian Aid appeal]. Neither actual nor generated titles are guaranteed to be unique across the corpus.

Where a monographic source text is divided into analytic texts, and different agencies control permissions for each analytic text, an id attribute is given for each <titlePiece>, so as to allow reconciliation with <list> elements. (See publicationStmt, §3.4.)

4 <author> <author>

Corpus header Not used

Text header Occurs once for each known author of an analytic text. Authors may be individual or corporate.

This and the following elements (<editor>, <imprint>, <idno> and <note>) are used only where the information for an analytic text differs from that for the monographic text of which it is a part.

4 <editor> <editor>

Corpus header Not used

Text header Occurs once for each known editor of an analytic text.

4 <imprint> <imprint>

Corpus header Not used

Text header Subordinate tags describe the publisher or originator of an analytic text.

5 <place> <place>

Corpus header Not used

Text header City or town in which the work was published or from which it is distributed.

5 <address> <address>

Corpus header Not used

Text header Not used (even if we have the address of the publisher, distributor, or release authority on file).

5 <publisher> <publishr>

Corpus header Not used

Text header A name, either individual or corporate. In the case of a published text, the name of a publisher; in the case of an unpublished text, the name of the distributor or release authority.

5 <date> <date>

Corpus header Not used

Text header In the case of a published analytic text, the date of publication; in the case of an unpublished text, the date of origination. Dates, which are given as accurately as possible, may vary in precision from an exact day to an approximate year, and are represented in accordance with [12] — for example, 1990-12-31, 1990-12, or 1990. *Circa* indicates an approximation, as in Circa 1990.

4 <idno> <idno>

Corpus header Not used

Text header An identifying number or code associated with an individual analytic text. Many occur more than once. The number or code is given as the value of the `n` attribute; the type or series of the number or code as the value of the `type`. (See notes below.) Examples of identifying numbers are the originator's filing references on unpublished or ephemeral items (`type=filing`) and the citation references given for papers in some journals (`type=citref`).

The element is empty.

4 <note> <note>

Corpus header Not used

Text header Gives additional information specific to an individual analytic text. May occur more than once. An example of such information might be details of the construction or pagination of a leaflet: `A4 sheet folded twice to give six printed sides`.

4 <title> <title>

Corpus header Not used

Text header The title of a monographic text. This element is always present. Where a monographic text has no clear title — for example, if it is made up of a collection of leaflets — a title will be generated by BNC staff. Square brackets will be used to distinguish such titles, as in `[Official documents 1]`. Where a text's actual title is used, this may not be unique across the corpus; generated titles, where used, will be unique.

4 <author> <author>

Corpus header Not used

Text header An author of a monographic text. Repeated for each known author. Authors may be individual or corporate.

Where a monographic text embodies one or more analytic texts, this and the following elements (<editor>, <imprint>, <idno> and <note>) are used only to give information which applies to all the analytic texts.

Where a monographic text is an issue of a serial work, the following elements are used to give information which applies only to the individual issue; information applying to the series as a whole is given by the elements following `title.series` — see below.

4 <editor> <editor>

Corpus header Not used

Text header An editor of a monographic text. Repeated for each known editor.

4 <imprint> <imprint>

Corpus header Not used

Text header Subordinate tags describe the publisher or originator of a monographic text, and are used as described for analytic texts above.

Where the monographic text is an element of a serial, only the <date> element of <imprint> is given, as this is specific to the monographic text. Information applying to all issues if the serial (`<place>` and `<publisher>`) is given by the `<imprint>` element associated with `<title.series>`. (See below.)

4 `<idno>` `<idno>`

Corpus header Not used

Text header An identifying number or code associated with a monographic text. May occur more than once. The number or code is given as the value of the `n` attribute; the type or series of the number or code as the value of the `type`. (See notes below.) Examples of identifying numbers are the International Standard Book Number (`type=isbn`) (see [11]) and the British Library Cataloguing In Publication reference (`type=blcip`).

The element is empty.

4 `<note>` `<note>`

Corpus header Not used

Text header Gives additional information specific to the monographic text. May occur more than once. An example of such information might be edition information for a book — `Second revised edition. First edition published 1975` — or of a particular issue of a serial work — `Late City edition`.

4 `<citnScope>` `<citnScope>`

Corpus header Not used

Text header This element is used only where material present in the original is absent in the electronic version for reasons not covered by `<editorialDecl>` (§4.3). Typically, this occurs where a sample of a text, rather than the whole text, has been captured. (The sampling procedure is described by the sampling declaration, §4.2.) The purpose of the element is both to define the extent of the sample taken from the source, and to give an indication of the full extent of the source. Absence of the element implies that, editorial procedures apart, the whole source text has been captured.

When present, the element gives, as the values of attributes, the number of the first page eligible for inclusion in a sample; the last page eligible; the first page actually included; and the last page actually included. The corresponding attribute names are `begSrc`, `endSrc`, `begSamp` and `endSamp`. The first and last eligible pages will normally be the start and end pages of the body text of the source text; however, prefatory pages may be included if some or all of the prefatory material appears in the electronic text. Similarly, appended pages may be included if some or all of a text's back matter has been captured.

It is generally acceptable for the values of `begSrc` and `endSrc` to be taken from different series of numbers: for example, the start page may be numbered *iii*, the end page *189*. There is no requirement that the endpoint of the first series of numbers (for example, *ix*) be given. In unusual cases, such as a manual in which pages in each section are numbered starting from one, an explanation may be given as the content of the element. In normal cases, no content need appear.

5 `<title.series>` `<title.s>`

Corpus header The title of the serial work (if any) of which the monographic work is a part. This element and those which follow (<author>, <editor>, <imprint>, <idno> and <note>) do not appear unless the monographic work is an element of a serial.

Text header

4 `<author>` `<author>`

Corpus header Not used

Text header An author of all issues of a serial. (Authors of individual issues are recorded in conjunction with <title> above.) Repeated for each known author. Authors may be individual or corporate.

This and the following elements (<editor>, <imprint>, <idno> and <note>) are used only to give information which applies to all the issues of a serial (or at least to those issues which BNC staff have to hand).

4 `<editor>` `<editor>`

Corpus header Not used

Text header An editor of all issues of a serial. Repeated for each known editor.

4 `<imprint>` `<imprint>`

Corpus header Not used

Text header Subordinate tags describe the publisher or originator of a serial, and are used as described for analytic texts above, except that the <date> element is not used.

4 `<idno>` `<idno>`

Corpus header Not used

Text header An identifying number or code associated with a serial. May occur more than once. The number or code is given as the value of the `n` attribute; the type or series of the number or code as the value of the `type`. (See notes below.) An example of identifying number is the International Standard Serial Number (`type=issn`). (See [9].)

The element is empty.

4 `<note>` `<note>`

Corpus header Not used

Text header Additional information pertaining to the all issues of a serial work. For example: `Published monthly. Distributed to members of the Campaign for Real Ale.`

3 `<citn.full>` `<cit.full>`

Corpus header Not used

Text header Not used

3 `<list.citn>` `<list.cit>`

Corpus header Not used

Text header Not used

Notes [13, II:23.5] provides three classes of citation. While <citn.full> provides the greatest capacity and flexibility, it is over-complex in most of the cases that the BNC must handle. Consequently, the BNC uses <citn.struct>. However, relative to [13], the content model for the element has been amended as set out in [6]. The resulting content model is:

```
<!ELEMENT citn.struct - o (((title.piece,
                            (author | editor | idno)*,
                            imprint?, note* )*,
                            title,
                            (author | editor | idno)*,
                            imprint?, citn.scope?, note*,
                            (title.series,
                            (author | editor | idno)*,
                            imprint?, note* )?           >
```

The first six elements of <citn.struct> (<author>, <titlePiece>, <editor>, <idno>, <imprint> and <note>) are repeated as necessary to describe each analytic work in a corpus text — for example, newspaper stories, papers from the proceedings of a conference, or leaflets collected into a single corpus text. These elements are used only where a text contains one or more analytic work. (A citation may list a single analytic work in a mongraphic work if, for example, the monographic work is a book of short stories, and just one story is captured for the BNC. Alternatively, the monographic work may be a particular issue of a learned journal which is given over to just one paper.)

The middle elements of <citn.struct> (<title>, <author>, <editor>, <idno>, <citn.scope> and <note>) describe the monographic work used as the source for the corpus text — for example, a book, an edition of a newspaper, the proceedings of a particular conference, or a collection of leaflets. Of these elements, at least <title> is always present.

The final elements of <citn.struct> (<title.series>, <author>, <editor>, <idno> and <note>), describe the serial work of which the monographic work used as the source for the corpus text is a part — for example, a newspaper, a series of conference proceedings, or a series of books. These elements are used only where a text is derived from a serial work.

Some examples should help to illustrate the use of <citn.struct> in the BNC:

```
<citn.struct>
  <title.piece>
    TELEVISION / Broken English: Mark Lawson
    on Home Run, and BBC2's Red Dynasty
  </title.piece>
  <author>Mark Lawson<author>
  <title.piece>
    DANCE / Taking a rain check: Judith Mackrell
    on Burrows and Smith at The Place, and The
    Royal Ballet's La Bayadere
  </title.piece>
  <author>Judith Mackrell<author>
  ...
  <title>
    &lsqb;The Independent, 1989-10-02, Arts section&rsqb;
  </title>
  <imprint>
```

```
    <date>1989-10-02</date>
  </imprint>
  <title.series>
    The Independent
  </title.series>
  <editor>Andreas Wittham-Smith</editor>
  <idno type=issn n='0999 9999'>
  <imprint>
    <place>London</place>
    <publisher>Independent Newspapers plc</publisher>
    <date>1989-10-02</date>
  <imprint>
</citn.struct>

<citn.struct>
  <title.piece>
    The Flight from Mind
  </title.piece>
  <author>Howard Robinson<author>
  <title.piece>
    &bquo;In the Beginning was the Deed&equo;;
    Mental development and the Philosophy of Mind
  </title.piece>
  <author>James Russell<author>
  ...
  <title>
    The Pursuit of Mind
  </title>
  <editor>Raymond Tallis</editor>
  <editor>Howard Robinson</editor>
  <idno type=isbn n=0 85635 918 1>
  <imprint>
    <place>Manchester</place>
    <publisher>Carcanet Press Limited</publisher>
    <date>1991</date>
  <imprint>
</citn.struct>

<citn.struct>
  <title.piece>
   Arranging a Funeral
  </title.piece>
  <title.piece>
    Help with Heating
  </title.piece>
  ...
  <title>
    &lsqb;Age Concern factsheets&rsqb;
  </title>
  <imprint>
    <place>London</place>
    <publisher>Age Concern</publisher>
    <date>Circa 1991</date>
  <imprint>
```

```
</citn.struct>

<citn.struct>
  <title.piece>
    Ethiopia The Million Tree Appeal
  </title.piece>
  <author>Jeff Thindwa</author>
  <imprint>
    <place>Northampton</place>
    <publisher>World Vision of Britain</publisher>
    <date>1990-03</date>
  </imprint>
  </title.piece>
  <title.piece>
    &lsqb;Oral Rehydration Therapy&rsqb;
  </title.piece>
  <author>Frances Carroll</author>
  <imprint>
    <place>London</place>
    <publisher>care Britain</publisher>
    <date>Circa 1990</date>
  </imprint>
  <title.piece>
  ...
  <title>
    &lsqb;Charitable appeals 1&rsqb;
  </title>
</citn.struct>

<citn.struct>
  <title>
    Nudists May Be Encountered
  </title>
  <author>Mary Scott</author>
  <idno type=isbn n='1 85242 173 8'>
  <idno type=lccip n='90 60285'>
  <imprint>
    <place>London</place>
    <publisher>Serpent's Tail</publisher>
    <date>1991</date>
  <imprint>
</citn.struct>
```

The recording of information about analytic works and serials is considered less important than recording information about monographic works. This means that corpus texts may initially be catalogued as monographic works, with details about any analytic parts, and serials to which the monographic works belong, being added later.

Information about overall authors and editors of serial works is collected on a best efforts basis: for example, we are unlikely to try to identify the commissioning editor of a series of books in cases where the editor is not named in the books themselves.

In general, books will not be catalogued as elements of a series unless they proclaim themselves to be such on their title page or its obverse. Thus, *The Best* *man to Die* by Ruth Rendell would be catalogued as a free-standing work, even though it is part of the author's *Inspector Wexford* series.

3.9 The Script Statement

3 `<scriptStmt>` `<scrpStmt>`

Corpus header Not used

Text header Not used for written texts. Used for spoken texts only if scripted, in which case it gives as full a citation as is practicable for the script or scripts used.

4 `<citn.struct>` `<cit.str>`

Corpus header Not used

Text header Used as described in §3.8. (Level numbers increase by one.) Where the text is taken from a single part of a serial work (for example, a cycle of plays), `<title>` gives the name of the series and a `<titlePiece>` element names the individual script represented in the text.

Where the script statement relates to a performance using multiple scripts (as in a current affairs programme, or a variety show), `<title>` gives the name of the whole performance and a `<titlePiece>` appears for each script. Where a performance is part of a series, `<title>` names the individual performance, and `<title.series>` the series. A (manufactured) example shows both of these cases:

```
<citn.struct>
  <title.piece>&lsqb;Tigray&rsqb;</titlePiece>
  <title.piece>&lsqb;Iraq&rsqb;</titlePiece>
  <title.piece>&lsqb;Bosnia&rsqb;</titlePiece>
  ...
  <title>
    &lsqb;World Service News,
    00:00 GMT 1991-09-01&rsqb;
  </title>
  ...
  <imprint>
    <date>1991-09-01</date>
  </imprint>
  <title.series>World Service News</title.series>
  ...
  <imprint>
    <place>London</place>
    <publisher>BBC World Service</publisher>
  </imprint>
...
</citn.struct>
```

(`<author>`, `<editor>`... elements omitted.)

Note that this is a citation for the scripts used in a broadcast, not for the broadcast itself. See `<broadcast>` in §3.10 for the means of citing a broadcast.

Script information will be given where possible for broadcast or theatrical performances recorded as part of the context-governed spoken corpus. Script information is unlikely to be recorded for matter read out by demographic spoken corpus participants (for example, books read to children, notices read at meetings etc.) (See also [4].)

3.10 The Recording Statement

3 `<recordingStmt>` `<recStmt>`

Corpus header Not used

Text header Describes the audio recording of a spoken text.

4 `<recording>` `<rec>`

Corpus header Not used

Text header Subordinate elements describe the audio recording associated with a spoken text. The `type` attribute is used only in the event that the recording is a video recording, in which case it has the value `video`, rather than the default `audio`. The TEI `dur` attribute is not used. The element occurs more than once in the (unlikely) event that a single text is derived from multiple recordings for which the contents of subordinate elements differ.

5 `<resp>` `<resp>`

Corpus header Not used

Text header Occurs once, giving rôle and name in each case, for each person or organization associated with the creation or subsequent editing of a recording. (May be a `<participant>` — see §7.2.)

6 `<name>` `<name>`

Corpus header Not used

Text header See `<resp>`

6 `<role>` `<role>`

Corpus header Not used

Text header See `<resp>`

5 `<equipment>` `<equipt>`

Corpus header Not used

Text header Briefly describes the equipment used to record a spoken text (if known).

5 `<broadcast>` `<broadcst>`

Corpus header Not used

Text header Cites the broadcast (if any) from which a spoken text is derived. Repeated in the unlikely event that a single text is derived from multiple broadcasts.

6 `<citn.struct>` `<cit.str>`

Corpus header Not used

Text header Used as described in §3.8. (See also example below.)

Notes An example of a citation of a broadcast (here an edition of an imagined series produced by an independent production company and televised by a regional broadcaster) might be:

```
<citn.struct>
  <title id=t01701>
    &rsqb;As it happens, 19:30 1992-11-05&lsqb;
  </title>
  <editor>Jocelyn Amhurst</editor>
  <note>Producer: William Smythe</note>
  <imprint>
    <place>Birmingham</place>
    <publisher>Central Television plc</publisher>
    <date>1992-11-05</date>
  <imprint>
  <title.series id=ts00082>As it happens</title.series>
  <editor>Anne Fawcett</editor>
  <imprint>
    <place>Bromsgrove</place>
    <publisher>Duff Diversions Ltd.</publisher>
  </imprint>
  <note>Producer: James Duff</note>
</citn.struct>
```

(Notice the use of `id` attributes: these may be necessary for the reconciliation of permissions information — see `<publicationStmt>`, §3.4.) The citation only describes a broadcast. Where a broadcast is scripted, a citation for the script or scripts should be given in `<scriptStmt>` (§3.9.

In the absence of some general means of consolidating information applying to many individual texts — but not all texts — into the corpus header, it seems likely that many texts will carry the same information about `<equipment>` in their headers. Entity references will be used to reduce the effect of this on the size of individual headers.

Citations will not be given for broadcasts which are incidental to other interactions — for example, a television programme which can be heard during a conversation.

4 The Encoding Description

2 `<encodingDesc>` `<encDesc>`

Corpus header Subordinate elements give encoding information relating to the corpus as a whole.

Text header Subordinate elements give encoding information relating to an individual text.

Notes See `<projectDesc>` (§4.1), `<samplingDecl>` (§4.2), `<editorialDecl>` (§4.3), `<refsDecl>` (§4.5), `<tagUsage>` (§4.4), and `<classDecl>` (§4.6) for subordinate elements.

Rather than repeat a lot of information about, for example, editorial practices, in the headers of many texts, it would be desirable to collect all such information together in the corpus header, and simply to reference the relevant sections from within the encoding descriptions of individual texts. Since it seems that [13] does not allow this, identical subsets of encoding information will appear in many text headers. Judicious use of entity references (one entity defined for each type of normalization, for example) will minimize the size of the text headers.

Note that, relative to [13], one new subordinate element, `<tagUsage>` (see §4.4) has been added.

4.1 The Project Description

3 `<projectDesc>` `<projDesc>`

Corpus header Prose description of BNC project

Text header Very brief description of the BNC project, together with an injunction to see the corpus header for fuller details.

4.2 The Sampling Declaration

3 `<samplingDecl>` `<sampDecl>`

Corpus header Prose description of the sampling procedures used for the BNC.

Text header `See corpus header`

Notes The sampling procedure being described here is that which selects less than the whole of a given source text for use in the corresponding electronic text. In the case of the BNC, this means the procedure by which 40,000 or fewer words are selected from the beginning, middle, or end of a text. Whether an electronic text contains the whole of an source text or one of the three types of sample is shown by means of its text classification — see §4.6 and §5.3.

4.3 The Editorial Practices Declaration

3 `<editorialDecl>` `<editDecl>`

Corpus header Subordinate tags describe each editorial practice common to all texts in the corpus.

Text header Subordinate tags describe editorial practices which pertain to a particular text, and which are not common to all texts in the corpus. The editorial practices used for a given text are the union of those given in the corpus header and those given in the text's header.

4 `<correction>` `<correctn>`

Corpus header Description of a correction method common to all corpus texts. The `status` attribute indicates the degree of correction applied (`high`, `medium`, `low` or `unknown`); the `method` attribute states how the correction has been applied (`silently` or with `tags`). The element may occur multiple times.

Text header Used as in the corpus header, but to describe a correction method used in an individual text.

4 `<normalization>` `<normn>`

Corpus header Description of a normalization method common to all corpus texts. The `source` attribute gives a brief citation of the authority for the method; the `method` attribute states how the normalization has been applied (silently or with tags). The element may occur multiple times.

Text header Used as in the corpus header, but to describe a normalization method used in an individual text.

4 <quotation> <quotn>

Corpus header Description of a means of handling quotations common to all corpus texts. The marks attribute states whether none, some, or all of the quotation marks in the source text have been retained as content; the form attribute indicates the means by which quotation marks are indicated within the text — as retained data; as rendition attributes; in some standardized (std) form; in an inconsistent, non-standardized (nonstd) form; or in some unknown manner. The element may occur multiple times.

Text header Used as in the corpus header, but to describe a means of handling quotations used in an individual text.

4 <hyphenation> <hyphn>

Corpus header Description of a means of handling hyphenation common to all corpus texts. The eol tag indicates whether all, some, or none of the line-end hyphenation in the source text has been retained. The element may occur multiple times.

Text header Used as in the corpus header, but to describe a means of handling hyphenation used in an individual text.

4 <segmentation> <segn>

Corpus header Description of a segmentation method common to all corpus texts. The element may occur multiple times.

Text header Description of a segmentation method used in an individual text. The element may occur multiple times.

4 <stdVals> <stdVals>

Corpus header Description of a means of representing standard values common to all corpus texts. May occur multiple times. (Unlikely to be used.)

Text header Description of a means of representing standard values used in an individual text. May occur multiple times. (Unlikely to be used, as the BNC does not standardize values.)

4 <analysis> <analysis>

Corpus header Description of an analysis method common to all corpus texts. One of these elements describes the word-class tagging applied by Lancaster. The element may occur multiple times.

Text header Description of an analysis method used in an individual text. May occur multiple times. Likely to occur for spoken texts containing overlapping speech, and for any text to which Lancaster has applied analysis additional to word-class tagging.

4 — <deletion>

Corpus header Description of an editorial deletion practice common to all texts. The method attribute states how deletion has been applied (silently or with tags). The element may occur multiple times.

Text header Used as in the corpus header, but to describe an editorial deletion practice used in an individual text.

Notes The scheme outlined results in much duplication of editorial information across multiple text headers. Entity references should be used to stand in for common elements. See further the notes under `<encodingDesc>`, §4.

The `<deletion>` element, which is additional to those defined in [13], describes the rules followed in making editorial deletions from captured text for the BNC project. Examples of such deletions are:

- advertisements;
- front and back matter;
- footnotes and endnotes;
- figures, diagrams and tables;
- poems;
- lists;
- labels;
- captions;
- epigraphs;
- forms;
- predominantly numeric material (e.g. lists of racing results);
- non-sentential lists (e.g. lists of supplier names and addresses, weather forecasts);
- material which cannot be represented using the BNC's writing system definition (see `<langUsage>`, §5.2); and
- truncation of `<div`*n*`>`s due to sampling strategy.

Note, however, that the last of these should be documented under `<samplingDecl>`, §4.2, rather than here.

4.4 The Tag Usage Declaration

3 *(none)* `<tagUsage>`

Corpus header Not used

Text header Subordinate elements summarize usage of individual tags in a the text.

4 *(none)* `<tagCount>`

Corpus header Not used

Text header Occurs once for each type of tag used in a text (that is, between `<text>` and `</text>`). The `gi` attribute gives tagname, the `count` attribute a count of start tags, and the `empty` attribute a value of either `y` or `n`, according to whether the element is allowed to have content or not. Since the distribution format for the BNC does not allow the markup minimization (see [7]), the number of start tags will always equal the number of end tags for non-empty elements; for empty elements, the number of end tags will always be zero.

The <tagCount> element normally has no content; however, in exceptional circumstances, the content may describe the usage of a tag in a particular corpus text: for example, `Lists marked only in front matter; body may contain unmarked lists`. Where such a comment is made, the `who` attribute should be used to identify its author.

Notes Relative to [13], this element is an addition to the contents of the <encodingDesc> element. Its content model is:

```
<!ELEMENT tagUsage (tagCount+)                                >
<!ATTLIST tagUsage %a.global                                  >
<!ELEMENT tagCount (#PCDATA)                                  >
<!ATTLIST tagCount %a.global
                   gi         NAME     #REQUIRED
                   count      NUMBER   #REQUIRED
                   empty      (y|n)    #REQUIRED
                   who        CDATA    #IMPLIED      >
```

4.5 The Reference Scheme Declaration

3 <refsDecl> <refsDecl>

Corpus header Description of a hierarchical reference scheme. `doctype` attribute has value `CDIF`.

Text header Pointer to correct <refsDecl> in corpus header. `doctype` attribute has value `CDIF`.

4 <step> <step>

Corpus header See notes below.

Text header Not used

4 <state> <state>

Corpus header Not used (required only for milestone-based schemes).

Text header Not used

Notes The BNC supports a single, three level, reference system, which applies to both written and spoken texts. It is based on the <s>s into which the word-class tagging at Lancaster divides the texts. Its definition is:

```
<refsD id=BNC10refs>
<step gi=bnc  delim='-' att=n>
<step gi=cdif delim='-' att=n>
<step gi=s att=n>
</refsD>
```

The following questions and answers (apologies for the condescending tone) should illuminate the thinking behind the scheme.

Q. What is the function of a reference system?

A. To give corpus users a standard means of specifying a citation for any part of the corpus in such a manner that any user can find the same location.

Q. What is the smallest part of the corpus that users might want to cite in this way?

A. A word. However, the overhead of encoding reference information with each word in the corpus is unacceptably high, so the BNC makes no provision for references this precise.

Q. Given that references to individual words will not be supported, what is the smallest element in the corpus to which reference will be allowed?

A. The `<s>` unit, roughly corresponding to a sentence. This will provide access at a granularity similar to that provided by line number references in existing corpora.

Q. Won't the BNC support reference by line number?

A. No. The CDIF markup provides no means of labelling lines (except in the case of poetry), so there is nowhere in the corpus to put line number information.

Q. Even if line numbers are not embedded in text files, can't I just make references by specifying that an external filter should be used to pick out a particular line number?

A. No. The BNC project makes no guarantee that the lineation of a file is the same each time it is delivered: for example, it may change to accommodate the line length limitations of a particular target system or delivery path. Indeed, a text may even be split into multiple files to accommodate the limitations of a system or path. Also, you may wish to change the lineation of your copy because you have inserted additional information which makes the original lines very long. Thus, your copy of a text may have lineation which differs from somebody else's copy, and quoting line numbers cannot provide a reliable reference.

Q. How about reference by page number, then?

A. This is not supported, either. While many BNC texts use `<pb>` tags to mark the top of a new page in the printed source text, not all do. Even in those texts which use `<pb>`, some page breaks may be unmarked — for example those corresponding to pages without a number printed on them. Thus there is no guarantee that any given page number will be marked, making page numbers unreliable as a reference mechanism.

Q. So how do I reference a particular `<s>`?

A. The full reference consists of three fields, separated by dashes (minus signs). The first field is the name of the corpus, `BNC1.0`. This appears as the value of the `n` attribute of the `<bnc>` element which contains the whole corpus, and its value will only change if new revisions of the corpus or entirely new corpora appear. The second field is the five- or six-character name of the text — the value of the `n` attribute of the text's enclosing `<cdif>` element. (This value also appears as the value of the `n` attribute in the text header file description `<idno>` tag, and may also be the name of the computer file containing the text.) The third field is a five-digit number, padded if necessary with leading zeros, which matches the value of the `n` attribute of the target `<s>`. Thus `BNC1.0-SexDis-01077` is a reference to the 1,077th `<s>` in the text `SexDis`.

Q. Why all those leading zeros?

A. So that references can easily be sorted into order with a sorting utility program. Most such programs provide an ASCII sort by default (although they can usually be persuaded by suitable command-line options to do numeric sorts). Under an ASCII sort, `BNC1.0-SexDis-01077` sorts after `BNC1.0-SexDis-00977` — the intuitively correct ordering, whereas, without leading zeros, `BNC1.0-SexDis-1077` would sort beforeBNC1.0-SexDis-977.

Q. What if I want to reference something bigger than a single <s>?

A. You specify a range of <s>s, such as `BNC1.0-SexDis-00977 – BNC1.0-SexDis-01077`.

Q. What if I want to reference the whole content of — say — a particular <div1>?

A. You specify its starting and ending <s>s.

Q. Why can't I just reference the <div1> itself?

A. The BNC's policy is to provide just one reference system. If there were additional reference systems for <div*n*>s (or any other element) as well as for <s>s, there would be more than one way to specify a reference for the same thing. For example, you could specify a <div1> by reference to the <div1> itself, or to the range of <s>s that it contains. There would be no way — other than examining the corpus itself — to tell that the two different types of reference referred to the same data. Such variability would defeat attempts to compare the findings of researchers on particular aspects of the corpus, and consequently the BNC does not provide for it.

Q. But the `n` attribute is used on <div*n*>s. Why can't I reference it?

SGML-aware software can certainly take notice of attribute values, and of the element nesting structure of a particular text. You can use the structure and the value of any attribute provided in the BNC— including `n` attributes on <div*n*>s — in any way that you like, *except* as a means of providing a reference used in citation. So as to be consistent with the references generated by other users of the corpus, you must only the <s>-based mechanism provided for this specific purpose. As a secondary issue, most texts do not provide <n> attribute values for some or all levels of <div>s, and those that do, take them verbatim from the original text. Thus, the first <div1> from an end sample with Roman numbering for its chapters might have `n=IX`. The same text might contain several <div2>s with `n=1`. Another text might start at `<div1 n=ONE>`. Taken together, these factors would make the use of <div> numbers an unreliable and inconsistent reference mechanism.

Q. How about utterances (<u>s) in a spoken text. Can I use them in a citation?

A. No. The same applies. You must reference the <s>s which make up the <u>.

Q. So I'm stuck with just referencing <s>s in all cases?

A. I'm afraid so.

4.6 The Classification Declaration

3 `<classDecl>` `<clasDecl>`

Corpus header Subordinate tags give definitions of the classification schemes used in the BNC.

Text header Not used

4 `<taxonomy>` `<taxonomy>`

Corpus header Occurs once for each classification taxonomy.

Text header Not used

5 `<citn>` `<citn>`

Corpus header Not used

Text header Not used

5 `<citn.struct>` `<cit.str>`

Corpus header If a citation is needed in connection with a classification scheme, used as described in §3.8. (Level numbers increase by three.)

Text header Not used

5 `<citn.full>` `<cit.full>`

Corpus header Not used.

Text header Not used

5 `<category>` `<category>`

Corpus header Where a classification scheme is not fully described by the citation given, this element appears once for each top-level classification feature (domain, time, author sex etc.).

Text header Not used

6 `<catdesc>` `<catdesc>`

Corpus header Describes the top-level selection or classification feature. For elements of the BNC written and spoken classification schemes described respectively in [2] and [3], the population being sampled and the method of sampling should be described.

Text header Not used

6 `<category>` `<category>`

Corpus header Appears once for each value which the selection or classification feature can take (male, female, mixed, unknown. . . .)

Text header Not used

7 `<catdesc>` `<catdesc>`

Corpus header Describes the value of the feature.

Text header Not used

Notes The scheme proposed in [15] can be accommodated through a citation of the Dewey Decimal classification scheme (or of [15] itself), possibly supplemented by a taxonomy having a single level of <category>s listing the classifications corresponding to each BNC domain.

As a means of representing the BNC's own classification scheme, the mechanism proposed is simple, having just two levels, but does all that is necessary **except** provide details of the relative proportions of each type of text within the corpus. In order to do this latter job, a much more complex nesting would be required, so that, for example, one could make it clear that the written part of the corpus constitutes 90% of the whole; imaginative texts constitute 20–30% of written texts; and works published between 1960 and 1974 make up 25% of imaginative texts. This is not considered worthwhile. (See also [5].)

A fragment of the proposed scheme would look like:

```
<classDecl>
  <taxonomy id=DeweyDec>
    <cit.struct>
      <!-- Citation for Dewey Decimal System -->
    </citn.struct>
  </taxonomy>
  <taxonomy id=BNCwrit>
    <citn.struct>
      <!-- Citation for BNCW08              -->
    </citStruct>
    <category>
      <catdesc>
        The date of first publication (published
        written works) or of origination
        (unpublished written works).
        ...
      </catdesc>
      <category id=date-1>
        <catdesc>
          1960&ndash;1974
        </catdesc>
      </category>
      <category id=date-2>
        <catdesc>
          1975&ndash;1993
        </catdesc>
      </category>
    </category>
    <category>
      <catdesc>
        The domain of a written work.
        ...
      <category id=domain-1>
        ...
```

5 The Profile Description

2 <profileDesc> <profDesc>

Corpus header Subordinate elements contain descriptions of the corpus in general.

Text header Subordinate elements contain descriptions of aspects of a particular text.

Notes See `<creation>` (§5.1), `<langUsage>` (§5.2), `<textClass>` (§5.3), `<textDesc>` (§7.1), `<particDesc>` (§7.2) and `<settingDesc>` (§7.4) for subordinate elements.

5.1 Creation

3 `<creation>` `<creation>`

Corpus header Not used

Text header Not used

Notes This TEI tag is intended to give details of the date and place of composition of a text. Since this information is rarely available for written texts in the BNC, the element cannot usefully be used for them; while the information is often available for spoken texts, it is recorded elsewhere in the header (see `<settingDesc>`, §7.4), and so need not be repeated here.

5.2 Language Usage

3 `<langUsage>` `<langUse>`

Corpus header Subordinate elements describe the language of the BNC and list languages known to be present in the BNC.

Text header Contains one `<language>` tag for each language known to be present in an individual text.

4 `<p>` `<p>`

Corpus header A series of paragraphs states that the main language of the BNC is modern British English, with exceptions in individual texts being noted in text headers, but not in general marked or quantified.

Text header Not used

4 `<language>` `<lang>`

Corpus header Repeated for each language known to be present in the BNC. The `n` attribute gives a standard name from [8] for the language[1]. Where [8] does not provide a name, other standards will be consulted[2], or BNC staff will generate a name. The optional `period` attribute (additional to those listed in [13]) describes the period from which the language usage dates. Values include `C12`, `C13` ... `C20` for the twelfth to twentieth centuries respectively. The `wsd` attribute references the writing system declaration used in the representation of the language. (This is the same in every case, since the BNC uses only one WSD and so may be defaulted in the DTD). The `usage` attribute is used as described in the notes below.

[1] [13] states that the `id` attribute should be used for this purpose. This seems to me to be an error.

[2] The ALLC may have something covering both current and non-current languages. More research is needed.

Text header Repeated for each language used in a particular text, with attributes used as described for the corpus header. Most texts are likely to have only `<language n=gbe wsd=BNCwsd usage=100> Modern British English</lang>` or similar.

Notes The BNC has no ambition to be a multi-lingual or dialectal corpus, and does not purport to be of use except as a corpus of modern British English. Given this design decision, the primary use of `<lang>` tags is to indicate that languages other than modern British English have been noted in a text during the proof-reading process, and may be a potential cause of bias in researchers' experimental results. For the purposes of this statement "modern British English" is defined as "British English as uttered at or after the earliest publication/production date allowed by the BNC sampling criteria." Thus, English dating from, say, 1900 (as quoted in a biography) is not modern by this definition. Clearly also, French, Latin, and other languages are excluded.

It should be noted that, even where multiple `<lang>` elements appear in the header of an individual text, the tagging of the text itself probably will not distinguish the language used at any point. (That is, the `lang` attribute will not be used in marking up the text.) This policy is a result of the cost of applying such markup, and may change if resources become available.

To date, fragments of many languages, and blocks of several modern Western European languages, Latin, and historical English and French have been encountered. Inevitably, texts also contain more or less modern American English. Best efforts are made to record the languages found in each text, and the approximate periods from which they date [3]. However, there is no guarantee that all languages used in a text will be noted, or that the details of those which are noted will be complete or correct.

The `usage` attribute is problematic, as it is defined by [13] to be a number. In cases where blocks of a language other than modern English occur, a proof-reader can give a finger-in-the-air figure of some multiple of 10% without appearing to be spuriously accurate. However, when there is a smattering of — say — old French, to state a figure of 2% is to to give the spurious impression that usage is neither 1% nor 3%. I therefore propose that the BNC tag corresponding to TEI's `usage` is `use`, and that it takes only the values `10%`, `20%` ... `100%`, `low` (to cover the case cited) and, as a default value, `remainder` (taken as being the difference between 100% and the sum of all the other `use` attributes).

A single writing system definition will be used for all corpus texts: text which cannot be represented using this WSD will be omitted. (See also [3]`<editorialDecl>`, §4.3.)

5.3 The Text Classification

3 `<textClass>` `<textClas>`

Corpus header Not used

Text header Subordinate tags give classification information for the text.

4 `<keywords>` `<keywords>`

Corpus header Not used

[3] It is generally possible to take a good guess at the period of a language fragment by reference to the immediate context

Text header Occurs at most once, listing keywords applying to text. The `scheme` attribute indicates source of keyword scheme used — see `<classDecl>` (§4.6). (Unlikely to be used, as keyword-based classification schemes are not likely to be used for the BNC.)

4 `<classCode>` `<clasCode>`

Corpus header Not used

Text header Gives a classification code for the text. The `scheme` attribute indicates source of classification scheme used — see `<classDecl>` (§4.6).

4 `<catRef>` `<catRef>`

Corpus header Not used

Text header `target` attribute references by ID the sampling classification categories to which the text belongs — see `<classDecl>` (§4.6).

Notes Given the `<category>` naming scheme shown in §4.6, an example of the use of the `<catRef>` tag, catering for both selection and classification features (see [2]), would be similar to

```
<catRef target='type-1 medium-1 select-2 domain-1
level-3 status-1 sample-1 time-1 author-1 autage-2
autet-00 autdo-18 autsex-2 aud-1'>
```

6 The Revision Description

2 `<revisionDesc>` `<revDesc>`

Corpus header Describes revisions to the corpus as a whole (addition of texts, modification of corpus header, systematic modification to all texts etc.).

Text header Describes revisions to an individual text.

3 `<change>` `<change>`

Corpus header Subordinate elements detail change. The `n` attribute gives the BNC version number in which the change first appeared. (Thus, `n=1.3` shows a change applied between versions 1.2 and 1.3.) (See `<editionStmt>`, §3.2.)

Text header Subordinate elements detail change. The `n` attribute gives the text revision number in which the change first appeared.

4 `<date>` `<date>`

Corpus header Date of change

Text header Date of change

4 `<by>` `<by>`

Corpus header Initials of institution responsible for change.

Text header Initials of institution responsible for change.

4 `<what>` `<what>`

Corpus header Description of the change.

Text header Description of the change.

Notes The <change> element and its contents are repeated for each recorded change. The most recent change is listed first. Hopefully, it will be possible to generate much of the revision description information for a text by automatic processing of the corresponding Z_ file. (See [1].)

I suggest that, at a minimum, the initials of the organization responsible for each change are recorded. While there seems to be some reluctance to identify individuals, particularly among the commercial participants in the BNC project, it would be useful to identify them, whether by their actual names, their initials, or anonymous identifiers. Data capture, proof-reading and post-editing styles are likely to vary from one person to another, and so affect document content. As I can imagine that future researchers would welcome information which could help them to take cognisance of such bias, consortium members are encouraged to document, with attribution, any significant change made to any text at any stage of the "sausage machine". That said, no decision has been taken on the amount of change information to be present in the BNC at the time of its initial release.

7 Optional Elements of the Structured Header

The following material, describing optional parts of the profile description (see §5, appears in [13, V:51 (maybe)]. However, since it is relevant to the CDIF header, it is discussed here. The information presented is based on [13, draft of 2 Apr 1992].

7.1 The Text Description

3 <textDesc> <textDesc>

Corpus header Subordinate elements describe usage in the corpus as a whole. (Or may be omitted — see notes.)

Text header Subordinate elements describe an individual text. (Or may be omitted — see notes.)

4 <channel> <channel>

Corpus header Description of the use of the <channel> tag in corpus texts

Text header Channel information for text. mode attribute (written, spoken, written to be spoken etc.) and information about the medium (print, face-to-face, manuscript. . .) can be derived automatically from BNC sampling criteria in most cases.

4 <constitution> <constitn>

Corpus header Description of the use of the <constitution> tag in corpus texts

Text header Constitution information for the text. type attribute (single, frags, composite, sample. . .) can be derived from sampling criteria and/or captured during proof-reading.

4 <derivation> <derivn>

Corpus header Description of the use of the <derivation> tag in corpus texts

Text header not stated

4 <domain> <domain>

Corpus header Description of the use of the <domain> tag in corpus texts

Text header Domain information for text. Values for type attribute (art, education, religious. . .) can be approximated by mapping from sampling criteria.

4 <factuality> <facty>

Corpus header Description of the use of the <factuality> tag in corpus texts

Text header Description of the factuality of the text. The type attribute (fact, fiction, unknown. . .) can fairly safely be automatically generated for books by a mapping from the sampling criteria. Other material is tagged as unknown.

4 <interaction> <interact>

Corpus header Description of the use of the <interaction> tag in corpus texts

Text header Description of interaction mediated by text. Pro formas (such as type=none from=one to=world) could cover many written texts; information for spoken texts could be determined only by examining the text, and in any event may well vary within an individual text. Consequently, the information is not likely to be captured.

4 <preparedness> <prepness>

Corpus header Description of the use of the <preparedness> tag in corpus texts

Text header Description of the preparedness of the text. Not stated in almost every case — unless it proves safe to tag all demographic spoken material as unprepared.

4 <purpose> <purpose>

Corpus header Description of the use of the <purpose> tag in corpus texts

Text header Description of the purpose of the text. Not stated in every case.

Notes The contents of the TEI <textDesc> tag are intended to provide a uniform means of representing "independent of an a priori theory of text-types" those aspects of a text which tend to be used in sampling criteria. The BNC has its own sampling criteria, implicitly based on a particular theory of text types. These criteria are represented in the <textClass> (§5.3). In some cases — for example, <channel> and <domain> — there is a mechanical mapping between the BNC sampling criteria and <textDesc> objects; in others there is not: the sampling criteria have nothing to say about, for example, <purpose>.

Since considerable effort could well be involved in creating a comprehensive <textDesc> for each text, I propose that this area of the text header is filled out only on a "best efforts" basis. This will probably mean that the mechanical stuff gets done, but that any tags requiring additional research into or examination of a text will simply contain some default value such as not stated. It may also mean that nothing is done, and that, consequently, <textDesc> does not occur in any BNC header. However, in the hope that at least some work can be done in this area, the element should be defined as an optional part of the CDIF header.

7.2 The Participants Description

3 `<particDesc>` `<partDesc>`

Corpus header Not used

Text header Description of the participants in the interaction mediated by a text.

4 `<participant>` `<partic>`

Corpus header Not used

Text header Written texts: not used.

Spoken texts: occurs once for each individual participant speaking in the text. Does not occur for non-speakers mentioned in the text. See 7.3 for content.

In all cases, `role`, `gender`, and `age` attributes are given.

4 `<particGroup>` `<particG>`

Corpus header Not used

Text header Written texts: not used.

Spoken texts: Occurs once for each group speaking in the text. (For example, a congregation.) See 7.3 for content.

In all cases, `role`, `gender` and `age` attributes are given.

4 `<particRelations>` `<particR>`

Corpus header Not used

Text header Written texts: not used.

Spoken texts: `<relation>` given for each recorded relationship of the participant.

5 `<relation>` `<relation>`

Corpus header Not used

Text header Spoken texts: describes relationship using `class`, `active`, `passive` and `mutual` tags. (From *relationship to respondent* question to demographic corpus respondents and *relationship to yourself* box in conversation log.)

Notes In this and subsequent sections, reference is made to [14], which contains samples of materials used for data capture in connection with the demographic spoken corpus.

As yet, no means is specified for tying together multiple spoken texts involving a given participant, or multiple written texts from a single author. This should be the subject of future work.

We need some standardized values for the `role` attribute, and for approximate values of `age` when exact values are not known. No proposals are being made at this stage.

7.3 The Demographic Description of a Participant or Group

5 <name> <name>

Corpus header Not used

Text header The full name of the participant, unless withheld for reasons of confidentiality, in which case the anonymous identifier used in the text is given.

5 — <status>

Corpus header Not used

Text header Used only if the marital and/or employment status of the participant is known. The optional marital attribute has the value s for a single participant, m for a married (or co-habiting) participant, d for a divorced, separated or widowed participant, or u (the default) if the participant's marital status is unknown. (From *Which of these describes you?* question to demographic respondents.) The optional employed attribute has the value y if the participant is in full- or part-time employment, n if unemployed (whether seeking work or not), or u if the participant's employment status is unknown. (From *Does respondent have a paid job full- or part-time?* question to demographic participants.) Although content is allowed, the element is normally empty.

5 <birthDate> <birthDt>

Corpus header Not used

Text header Used only if at least the year of birth of the participant is known. (From *Age last birthday* questions for demographic participants — see [14].) (See also age attribute of <participant> and <particGroup> in §7.2.)

5 <birthPlace> <birthPl>

Corpus header Not used

Text header Used only if the birth-place of the participant is known. (From *in which town were you born* question for demographic respondents only.) Unlikely to be given for groups — although could conceivably be useful for indicating location of headquarters of a corporation.

5 — <pastRes>

Corpus header Not used

Text header Gives town and county for a past residence of the participant, if known. May occur multiple times. (From *in which town were you living when you were at primary/secondary school* and *In what other places have you lived for more than three years?* questions to demographic respondents.)

5 <firstLang> <firstLg>

Corpus header Not used

Text header Used only if the first language of the participant is known to us. Can occur more than once for participants brought up in a multilingual environment. (For demographic participants, the only information captured is that their first language is not English. We do not know what the first language is in these cases.)

5 `<langKnown>` `<lgKnown>`

Corpus header Not used

Text header Used only if we are aware of other languages known to the participant besides those listed under `<firstLang>`. Can occur more than once. List should include English if `<firstLang>` is not English. (Information not captured for demographic participants.)

5 — `<ethnic>`

Corpus header Not used

Text header Used only if we are aware of the ethnic origin of the participant. (From *to which of the ethnic groups shown do you consider you belong?* question to demographic respondents.)

5 — `<accent>`

Corpus header Not used

Text header Used only if or we are aware of a regional, class or foreign accent for the participant. Unlikely to be used for groups. (Taken from *Regional accent* question for demographic participants.)

5 `<residence>` `<residce>`

Corpus header Not used

Text header Used only if the place of residence of the participant at the time of the production of the text is known. (Town and county of sampling point for demographic respondents only.) The optional `hoh` attribute has the value `y` if the participant is known to be the head of the household; `n` if known not to be; or `u` (the default) if the status is unknown. (From *Establish who is the head of household* question to demographic respondents.) The optional `timeres` attribute gives the number of years and months (`yy-mm`)for which the participant has been resident. (From *How long have you been living at this address* question to demographic respondents.)

5 `<education>` `<educn>`

Corpus header Not used

Text header Used only if the educational attainments of the participant are known. Occurs at most once, describing the highest level of attainment. (From *at what age did you finish full-time education* question for demographic respondents only.) Where a participant is at school, and we have details of the type of school, this information is given. (From *Which types of school?* question to respondents about children in household.) Where we are aware of the place of places at which a participant attended schools or other educational establishments, and of any qualifications obtained, this information is given. (From *in which town were you living when you attended primary/secondary school* and *Do you have any qualifications?* questions to demographic respondents.)

5 `<affiliation>` `<affiln>`

Corpus header Not used

Text header Used only if the affiliation(s) of the participant are known. May occur more than once.

5 `<occupation>` `<occupn>`

Corpus header Not used

Text header Used only if the occupation of the participant is known. (From job-related questions for demographic respondents and members of their households). The optional `rank` attribute names the position, rank or grade in which the participant works, if known. (For example, `sergeant`, `manager`.) The optional `size` attribute gives an approximate number of employees at the participant's workplace; the optional `reports` attribute, the approximate number of people reporting to the participant. The element may occur more than once, giving several occupations. (e.g. parent, voluntary worker, school governor…) (However, this information is not captured from demographic corpus respondents.)

5 `<socecstatus>` `<socstat>`

Corpus header Not used

Text header Used only if the socio-economic status of the participant is known. Given through the value of the `name` attribute: one of `AB`, `C1`, `C2`, `DE`. (From *now assess grade of respondent's job* item for demographic participants only.) Tag should never have content.

5 — `<media>`

Corpus header Not used

Text header Lists the print, audio or visual media preferred by the participant, if known. Occurs once for each medium for which information is available. The required `type` attribute gives the type of medium. Sample values are `television`, `radio`, `'Sunday paper'`, `'daily paper'` and `book`. The optional `hours` attribute gives the number of hours per week spent with the medium. Sample values are `none`, `low`. `1`, `2` … `9` and `u`. The last (unknown) is the default.) (From *leisure time* questions to demographic participants.)

5 — `<hobbies>`

Corpus header Not used

Text header Lists the hobbies of the participant, if known. (From *hobbies or interests* question to demographic participants.)

Notes Note that none of the tags above are used for written texts. Several of the tags are additional to those suggested by [13], and are included so as to allow the recording of all the information captured about respondents and participants in the demographic spoken corpus. (See [14].)

We need a long but standardized list for birth-places and residences (north-west England, Scotland, India etc.), in which case the `name` attribute to `<birthPlace>` could be used to carry the information. ISO country codes from [10] can be used in most cases, although greater accuracy is needed inside the UK, and lower accuracy for generalizations such as "middle east".

A short standardized list could also be used for `<education>`. Following the format of the data captured for the demographic corpus question *at what age did you finish your full-time education?*, this would be `14 or under`, `15`, `16`, `17`, `18`, `19`, `'20 or over'` and `'still studying'`.

7.4 The Setting Description

3 `<settingDesc>` `<setDesc>`

Corpus header Description of the use of the `<settingDesc>` tag in individual texts.

Text header Written texts: not used (but see notes below).

Spoken texts: subordinate tags describe setting.

4 `<setting>` `<set>`

Corpus header Not used

Text header Spoken texts only: setting in which the interaction takes place, where this can be determined from notes in conversation record's *What were you doing while recording?* box (see [14, Appendix B]) or other information. Generally occurs only once, describing the setting of all participants, in which case the `who` attribute is not used. May occur more than once for, for example, telephone conversations or for radio broadcasts incidental to a spoken text. The `who` attribute is used in such cases to identify the participant(s) at a particular location. (Note, however, that information of this type is not captured for the demographic spoken corpus.)

5 `<place>` `<place>`

Corpus header Not used

Text header Spoken texts only: place where interaction takes place, if this can be determined from the conversation record question *In which town/city/village did the conversation take place?* question, or from other information. Includes the name of one of the three demographic sampling regions (fieldwork areas) given in [14, Appendix A] as the value of the `n` attribute. The content of the tag may give more specific information.

5 `<time>` `<time>`

Corpus header Not used

Text header Spoken texts only: date and time at which recording was made. The `recBeg` attribute gives, in [12] format (`YY-MM-DD HH:MM`), a date and time at or after which recording started; the `recEnd` attribute a time at or before which recording ended. If only `recBeg` is given, it specifies an exact time at which recording started. (See notes below).

5 `<locale>` `<locale>`

Corpus header Not used

Text header Spoken texts only: locale in which the interaction took place, if this can be determined from notes in the conversation records.

5 `<activity>` `<activity>`

Corpus header Not used

Text header Spoken texts only: activity during which the interaction took place, if this can be determined from notes in the conversation records.

Notes No setting information is recorded for written texts as, in most cases, any information that might be recorded concerning the producers and receivers of such texts would be entirely subjective and speculative. Exceptions could conceivably be made for specialized areas such as e-mail.

The information recorded in the respondents' conversation records for the demographic corpus includes the date and time at which recording commenced on each 45-minute tape segment. We can therefore say that an interaction recorded on a particular segment happened no earlier than the start time for the recording of that segment, and no later than the start time for the recording of the next segment, or, for the final segment recorded, of the final interview.

References

[1] Gavin Burnage, TGCW35, *Corpus Text Processing: Directory Structures and File Names*, August 1992

[2] Jeremy Clear, BNCW08, *Written Corpus Design Specification*, September 1991

[3] Crowdy et al, TGAW14, *Spoken Corpus design Specification*, October 1991

[4] Crowdy et al, TGCW21, *Spoken Corpus Transcription Guide*, December 1991

[5] Dunlop, TGCW20, *[Corpus text selection and classification features]*, December, 1991

[6] Dunlop, *[exchange of electronic mail with C. M. Sperberg-McQueen re citn.struct]*, August 1992

[7] Charles S. Goldfarb, *The SGML Handbook*, Oxford University Press, 1990

[8] ISO 639:1988, *Code for the representation of names of languages*, International Organization for Standardization

[9] ISO 2108:1978, *Documentation — International Standard Book Numbering (ISBN)*, International Organization for Standardization

[10] ISO 3166:1988, *Codes for the representation of names of countries*, International Organization for Standardization

[11] ISO 3297:1986, *Documentation — International Standard Serial Numbering (ISSN)*, International Organization for Standardization

[12] ISO 8601:1988, *Data elements and interchange formats — Information exchange — Representation of dates and times*, International Organization for Standardization

[13] C. M. Sperberg-McQueen & Lou Burnard (eds.), *Guidelines for the Encoding and Interchange of Electronic Texts — TEI P2*, Oxford and Chicago, 1992

[14] BMRB, BNCP25, *Longman Spoken Corpus Technical Report*, July 1992

[15] Russell Sweeny, TGAP19, *British National Corpus — Selection and Classification of Texts*, April 1992

[16] *TEI.1 toy DTD version 2*, June 1991