Reference Guide to BNC Baby (second edition)

edited by Lou Burnard

July 30, 2008

edited by Lou Burnard July 2008

Contents

1 Introduction

This manual describes BNC Baby, a four million word sampling of the 100 million word British National Corpus. It contains a brief description of the design of this sample and information about the way in which it is encoded. A list giving brief bibliographic details for each text making up the sample is also included.

The present document is derived from the Reference Guide for the British National Corpus (XML Edition) provided as a part of the BNC XML Edition, first released in March 2007, and available on the web. Definitive information about the design principles, sampling methods, and encoding principles of the BNC should be taken from that manual. Further information about the BNC is also available from its World Wide Web server at `http://www.natcorp.ox.ac.uk`

The BNC was originally created by an academic-industrial consortium whose original members were:

- Oxford University Press
- Longman Group Ltd
- Chambers Harrap
- Oxford University Computing Services (OUCS)
- Unit for Computer Research on the English Language (Lancaster University)
- British Library Research and Development Department

Creation of the corpus was funded by the UK Department of Trade and Industry and the Science and Engineering Research Council under grant number IED4/1/2184 (1991-1994), within the DTI/SERC Joint Framework for Information Technology. Additional funding was provided by the British Library and the British Academy.

After the completion of the first edition of the BNC, a phase of tagging improvement was undertaken at Lancaster University with funding from the Engineering and Physical Sciences Research Council (Research Grant No. GR/F 99847). This tagging enhancement project was led by Geoffrey Leech, Roger Garside and Tony McEnery. Correction and validation of the bibliographic and contextual information in all the BNC Headers was also carried out for this second version of the corpus, known as the BNC World Edition.

In March 2007, a third version of the corpus was made available in XML format, known as the BNC XML Edition. No new texts were added but additional information about head-words (lemma) and simplified word-class tags was included. The treatment of multi-word units and extra-linguistic information was revised, as detailed in Reference Guide for the British National Corpus (XML Edition).

BNC Baby was produced at OUCS by Lou Burnard, Martin Wynne, and Ylva Berglund. It was the first version of the BNC to be distributed entirely in XML. In preparing this second edition, some minor errors in the manual have been corrected.

2 Design of BNC Baby

BNC Baby is a four-part corpus constructed by a principled sampling of the BNC XML Edition. The principles underlying selection of texts for inclusion in BNC Baby may be summarized as follows:

- equal amounts of material (approximately one million words each) have been selected from each of four domains represented in the full BNC: academic writing, imaginative writing, newspaper texts, and spontaneous conversation.

- texts previously included in the BNC Sampler were excluded from selection.
- texts were copied directly from the BNC XML release (March 2008), without editing or revision.

To select materials meeting these criteria, the following procedure was adopted. Selection of texts was made on the basis of the information about them recorded in the original corpus, in particular the text classifications provided by David Lee (Lee, 2001).

2.1 Fiction

The BNC contains 477 texts (16 million words) classed as ‘imaginative’. These vary considerably in date of publication, target audience and medium, as well as in author properties and other features. Compared with other texts in the corpus, these texts are also rather long. Achieving a comparable variety in the one million word subset would be difficult, if not impossible. In BNC Baby therefore, we selected according to the following carefully defined criteria. Texts for the fiction component were selected from texts classified as "written imaginative", published as books between 1985-1994, as having been produced for an adult audience, and having the genre label W fict prose. From this set of 356 texts, a random sample of about one million words (25 texts) was drawn. The sample was checked to ensure no more than one title by any particular author was selected.

2.2 Newspapers

In the design of the original BNC, newspapers were not identified as a distinct category, although a very large amount of newspaper material is contained in it. On the basis of the descriptive information provided in the text headers, it is however possible to select different kinds of newspaper text. Note that each ‘text’ is made up of several newspaper articles, often drawn from a particular subject domain, rather than complete issues of a paper. In this, BNC Baby follows the sampling methods of the original corpus compilers.

The BNC Baby selection was made to ensure a mix of national and local papers, wide coverage of topics and little duplication of dates. Approximately 60% of the newspaper data comes from five national papers, and the remaining 40% from regional newspapers. An attempt was made to include texts from different domains and genres, to maximise the spread in topic areas covered. As far as possible, the texts were selected to ensure a spread of dates of publication in order to minimise the effects of seasonal or topical variation. The size of the texts was considered and choices made to ensure a roughly equal distribution across different newspapers within the national and regional subsets respectively, as well as a spread across subject areas. Differences between the amount of data from each newspaper in the component is largely due to the considerable variation in the size of the newspaper texts in the BNC.

The following table shows the number of words in each newspaper sampled for BNC Baby

Newspaper	words	%
Daily Mirror	124251	12%
Daily Telegraph	128794	13%
Guardian	129598	13%
Independent	131205	13%
Today	91238	9%
Belfast Telegraph	43006	4%
East Anglian Daily Times	43674	4%
Liverpool Daily Post and Echo	85441	9%
Northern Echo	68887	7%
The Alton Herald	56316	6%
The East Anglian	15814	2%
The Scotsman	66709	7%
Ulster Newsletter	16888	2%
Total component	1001821	100%

2.3 Spoken

The conversational data in BNC Baby has been drawn only from the spoken demographic component of the BNC. Each text consists of a number of conversations recorded by one individual, capturing data produced by a number of different speakers in different situations. Speakers were recruited (as described elsewhere) according to demographic principles, in order to be broadly representative of the UK population in terms of age, gender, region, and class.

Texts for which very little information about the speakers was available were excluded from selection. From the remainder, 30 texts were then randomly selected.

The following table[1] shows the number of words spoken by participants in the spoken part of BNC Baby, broken down by sex, age group, and social class:

Table 2: Speakers in BNC Baby (first edition)

	value	words	% of category	% of corpus
Age	0-14	102,350	11%	10%
	15-24	73,891	8%	7%
	35-44	182,976	20%	18%
	25-34	292,083	32%	29%
	45-59	113,038	12%	11%
	60+	159,948	17%	16%
	total	924,286	100%	91%
Sex	Female	551,077	59%	55%
	Male	384,337	41%	38%
	total	935,414	100%	93%
Class	AB	243,125	35%	24%
	C1	232,981	33%	23%
	C2	133,963	19%	13%
	DE	86,189	12%	9%
	total	696,258	100%	69%

2.4 Academic

In the design of the original BNC, academic prose is not identified as a distinct category, although a large amount of such material is contained in it. On the basis of the descriptive information provided in the text headers, it is however possible to select such texts. From the set of texts identified by David Lee as "written academic", titles were randomly selected within different subject areas to maximize variation in topic. An attempt was also made to include data originally published in periodicals as well as in books, although no targets were set for the proportions of material from each medium. Of the 501 academic writing texts in the BNC World Edition, 30 were selected for the BNC Baby academic component.

2.5 Design of the British National Corpus

Section 1 of the Reference Guide for the British National Corpus (XML Edition) gives a detailed overview of the design principles underlying the original construction of the British National Corpus, which principles are necessarily followed in the sampling for BNC Baby. The *Reference Guide* also includes detailed information about the actual composition of the full corpus with respect to its selection and classification, as well as detailed information about the way in which it is encoded, covering both the markup conventions applied and the linguistic annotation with which the corpus was enriched.

[1] These counts were calculated for the first edition of the BNC Baby and may not therefore exactly match the counts obtained for the second edition.

We do not duplicate that information here: it should however be consulted for a proper understanding of the composition of the BNC Baby corpus. We do however include some summary tables for comparative purposes.

The following table gives a more detailed breakdown of the genres represented in BNC Baby, using the classifications identified in Lee 2001.

Table 3: Genre classification for BNC Baby

Classification	Text	W-units	%	S-units	%
S conv	30	1017025	25.39	155921	46.82
W ac:humanities arts	7	224872	5.61	8578	2.57
W ac:medicine	2	89821	2.24	4120	1.23
W ac:nat science	6	215549	5.38	9441	2.83
W ac:polit law edu	6	195836	4.89	7993	2.40
W ac:soc science	7	209645	5.23	9206	2.76
W ac:tech engin	2	77533	1.93	3590	1.07
W fict prose	25	1010279	25.23	82002	24.62
W newsp brdsht nat: arts	9	36603	0.91	1748	0.52
W newsp brdsht nat: commerce	7	64162	1.60	3075	0.92
W newsp brdsht nat: editorial	1	8821	0.22	441	0.13
W newsp brdsht nat: misc	25	121194	3.02	6036	1.81
W newsp brdsht nat: report	3	48190	1.20	2342	0.70
W newsp brdsht nat: science	5	18245	0.45	958	0.28
W newsp brdsht nat: social	13	34516	0.86	1797	0.53
W newsp brdsht nat: sports	3	36796	0.91	1846	0.55
W newsp other: arts	3	43687	1.09	2504	0.75
W newsp other: commerce	5	89170	2.22	5051	1.51
W newsp other: report	7	232739	5.81	13045	3.91
W newsp other: science	7	13616	0.34	680	0.20
W newsp other: social	8	94676	2.36	4903	1.47
W newsp tabloid	1	121252	3.02	7686	2.30

For comparison, the following table shows the breakdown of texts making up BNC Baby with respect to the original domain classifications used in the full BNC:

Table 4: Written Domain

	texts	w-units	%	s-units	%
Imaginative	25	1010279	25.23	82002	24.62
Informative: natural & pure science	13	289022	7.21	12428	3.73
Informative: applied science	14	215402	5.37	10377	3.11
Informative: social science	20	378608	9.45	16392	4.92
Informative: world affairs	21	475132	11.86	23340	7.00
Informative: commerce & finance	12	153332	3.82	8126	2.44
Informative: arts	17	161818	4.04	7714	2.31
Informative: belief & thought	11	51010	1.27	2262	0.67
Informative: leisure	19	252599	6.30	14401	4.32

The following tables indicate the breakdown according to the selection criteria for spoken respondents. Note that the figures relate to respondents only, rather than all speakers recorded.

Age-group(cont.)

Table 5: Age-group

	texts	w-units	%	s-units	%
Respondent Age 0-14	5	39058	0.97	6181	1.85
Respondent Age 15-24	3	83532	2.08	15250	4.58
Respondent Age 25-34	7	321989	8.04	46652	14.01
Respondent Age 35-44	8	253862	6.33	40397	12.13
Respondent Age 45-59	3	170836	4.26	25736	7.72
Respondent Age 60+	4	147748	3.68	21705	6.51

Table 6: Social class

	texts	w-units	%	s-units	%
AB respondent	9	293513	7.33	46986	14.11
C1 respondent	9	248498	6.20	39727	11.93
C2 respondent	8	278139	6.94	36871	11.07
DE respondent	4	196875	4.91	32337	9.71

Table 7: Sex

	texts	w-units	%	s-units	%
Unknown	1	3189	0.07	455	0.13
Male respondent	12	338814	8.46	51383	15.43
Female respondent	17	675022	16.85	104083	31.25

3 Basic structure

The original British National Corpus was provided as an application of ISO 8879, the Standard Generalized Mark-Up Language (SGML). 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 legally be combined. SGML was a predecessor of XML, the extensible markup language defined by the World Wide Web Consortium and now in general use on the World Wide Web. XML was originally designed as a means of distributing SGML documents on the web.

This XML edition of the BNC is delivered in an XML format which is documented in this manual in section *3.1. Markup conventions* below; more detailed information about XML itself is readily available in many places.[2]

The original BNC encoding format was also strongly influenced by the proposals of the Text Encoding Initiative (TEI). This international research project resulted in the development of a set of comprehensive guidelines for the encoding and interchange of a wide range of electronic texts amongst researchers. An initial report appeared in 1991, and a substantially revised and expanded version in early 1994. A conscious attempt was made to conform to TEI recommendations, where these had already been formulated, but in the first version of the BNC there were a number of differences in tag names, and models. In the second edition of the BNC (BNC World), the tagging scheme was changed to conform as far as possible with the published *Recommendations* of the TEI (TEI P5). In the XML edition, this process has continued, and the corpus schema is now supplied in the form of a TEI customization, as further documented in the full reference manual. .

[2] The article in Wikipedia (`http://en.wikipedia.org/wiki/XML`) is probably as good a starting point as any; another is at `http://homepages.inf.ed.ac.uk/wadler/xml/`

3.1 Markup conventions

The BNC XML edition is marked up in XML and encoded in Unicode. These formats are now so pervasive as to need little explication here; for the sake of completeness however, we give a brief summary of their chief characteristics. We strongly recommend the use of XML-aware processing tools to process the corpus; see further *8. Software for BNC Baby*.

An XML document, such as the BNC consists of a single root element, within which are nested occurrences of other element types. All element occurrences are delimited by tags. There are two forms of tag, a start-tag, marking the beginning of an element, and an end-tag marking its end (in the case of 'empty elements', the two may be combined; see below). Tags are delimited by the characters < and >, and contain the name of the element (its gi, for generic identifier), preceded by a solidus (/) in the case of an end-tag.

For example, a heading or title in a written text will be preceded by a tag of the form `<head>` and followed by a tag in the form `</head>`. Everything between these two tags is regarded as the content of an element of type `<head>`.

Attributes applicable to element instances, if present, are also indicated within the start-tag, and take the form of an attribute name, an equals sign and the attribute value, in the form of a quoted literal. Attribute values are used for a variety of purposes, notably to represent the part of speech codes allocated to particular words by the CLAWS tagging scheme.

For example, the `<head>` element may take an attribute type which categorizes it in some way. A main heading will thus appear with a start tag `<head type="MAIN">`, and a subheading with a start tag `<head type="SUB">`.

The names of elements and attributes are case-significant, as are attribute values. The style adopted throughout the BNC scheme is to use lower-case letters for identifiers, unless they are derived from more than one word, in which case the first letter of the second and any subsequent word is capitalized: examples include `<teiHeader>` or `<particDesc>` (for 'participant description').

Unless it is empty, every occurrence of an element must have both a start-tag and an end-tag. Empty elements may use a special syntax in which start and end-tags are combined together: for example, the point at which a page break occurs in an original source is marked `<pb/>` rather than `<pb></pb>`

The BNC is delivered in UTF-8 encoding: this means that almost all characters in the corpus are represented directly by the appropriate Unicode character. The chief exceptions are the ampersand (&) which is always represented by the special string `&amp;`, the double quotation mark, which is sometimes represented by the special string `&quot;`, and the arithmetic less-than sign, which always appears as `&lt;`. These 'named entity references' use a syntactic convention of XML which is followed by this version of the corpus. All other characters, including accented letters such as é or special characters such as —, are represented directly.

The number of linebreaks in the corpus has been reduced to a minimum in order to simplify processing by non-XML aware utilities. In particular:

- XML tags are never broken across linebreaks;
- the TEI Header prefixed to each text contains no linebreaks
- each `<s>` element begins on a new line

. Many XML aware utilities are available to convert this representation as required.

3.2 An example

Here is the opening of text J10 (a novel by Michael Pearce). In this example, as elsewhere, we have placed each element on a separate line for clarity; this is not a requirement of XML however.

```
<wtext type="FICTION">
 <pb n="5"/>
 <div level="1">
```

```
  <head>
   <s n="1">
    <w c5="NN1" hw="chapter" pos="SUBST">CHAPTER </w>
    <w c5="CRD" hw="1" pos="ADJ">1</w>
   </s>
  </head>
  <p>
   <s n="2">
    <c c5="PUQ">‘</c>
    <w c5="CJC" hw="but" pos="CONJ">But</w>
    <c c5="PUN">,</c>
    <c c5="PUQ">’ </c>
    <w c5="VVD" hw="say" pos="VERB">said </w>
    <w c5="NP0" hw="owen" pos="SUBST">Owen</w>
    <c c5="PUN">,</c>
    <c c5="PUQ">‘</c>
    <w c5="AVQ" hw="where" pos="ADV">where </w>
    <w c5="VBZ" hw="be" pos="VERB">is </w>
    <w c5="AT0" hw="the" pos="ART">the </w>
    <w c5="NN1" hw="body" pos="SUBST">body</w>
    <c c5="PUN">?</c>
    <c c5="PUQ">’</c>
   </s>
  </p>
   ....</div>
</wtext>
```

This example has been reformatted to make its structure more apparent: as noted above, in the actual corpus texts, newlines appear only at the start of each <s> element, rather than (as here) at the start of each element. The original files also lack the extra white space at the start of each line, used in the above example to indicate how the XML elements nest within one another.

The example begins with the start tag for a <wtext> (written text) element, which bears a type attribute, the value of which is FICTION, the code used for texts derived from published fiction. The start tag is followed by an empty <pb> element, which provides the page number in the original source text. This in turn is followed by the start of a <div> element, which contains the first subdivision (chapter) of this text. This first chapter begins with a heading (marked by a <head> element) followed by a paragraph (marked by the <p> element). Further details and examples are provided for all of these elements and their functions elsewhere in this documentation.

Each distinct word and punctuation mark in the text, as identified by the CLAWS tagger, has been separately tagged with a <w> or <c> element as appropriate. These elements both bear a c5 attribute, which indicates the code from the CLAWS C5 tagset allocated to that word by the CLAWS POS-tagger; <w> elements also bear a pos attribute, which provides a less fine-grained part of speech classification for the word, and an hw attribute, which indicates the root form of the word. For example, the word ‘said’ in this example has the CLAWS 5 code VVD, the simplified POS tag VERB, and the headword say. The sequence of words and punctuation marks making up a complete segment is tagged as an <s> element, and bears an n attribute, which supplies its sequence number within the text. A combination of text identifier (the three letter code) and <s> number may be used to reference any part of the corpus: the example above contains J10 1 and J10 2.

This is not, of course, a complete text: in particular, it lacks the TEI header which is prefixed to each text file making up the corpus. Its purpose is to indicate how the corpus is encoded. Any XML aware processing software, including common Web browsers, should be able to operate directly on BNC texts in XML format.

The remainder of this manual describes in more detail the intended semantics for each of the XML elements used in the corpus, with examples of their use.

3.3 Corpus and text elements

The BNC contains a large number of text samples, some spoken and some written. Each such sample has some associated descriptive or bibliographic information particular to it, and there is also a large body of descriptive information which applies to the whole corpus.

In XML terms, the corpus consists of a single element, tagged `<bnc>`. This element contains a single `<teiHeader>` element, containing metadata which relates to the whole corpus, followed by a sequence of `<bncDoc>` elements. Each such `<bncDoc>` element contains its own `<teiHeader>`, containing metadata relating to that specific text, followed by either a `<wtext>` element (for written texts) or an `<stext>` element (for spoken texts).

Each bncDoc element also carries an xml:id attribute, which supplies its standard three-character identifier.

```
<bncDoc xml:id="ABC">
 <teiHeader>
<!- header for document ABC here ->
 </teiHeader>
 <wtext type="FICTION">
<!- sentences from ABC (a written text classified as fiction) here ->
 </wtext>
</bncDoc>
```

The components of the TEI header are fully documented in section *6. The header*.

Note that different elements are used for spoken and written texts because each has a different substructure; this represents a departure from TEI recommended practice.

3.4 Segments and words

The `<s>` element is the basic organizational principle for the whole corpus: every text, spoken or written, is represented as a sequence of `<s>` elements, possibly grouped into higher-level constructs, such as paragraphs or utterances. Each `<s>` element in turn contains `<w>` or `<c>` elements representing words and punctuation marks.

The n attribute is used to provide a sequential number for the `<s>` element to which it is attached. To identify any part of the corpus uniquely therefore, all that is needed is the three character text identifier (given as the value of the attribute xml:id on the `<bncDoc>` containing the text, followed by the value of the n attribute of the `<s>` element containing the passage to be identified.

These numbers are, as far as possible, preserved across versions of the corpus, to facilitate referencing. This implies that the sequence numbering may have gaps, where duplicate sequences or segmentation errors have been identified and removed from the corpus. In a few (about 700) cases, sequences formerly regarded as a single `<s>` have subsequently been split into two or more `<s>` units. For compatibility with previous versions of the corpus, the same number is retained for each new `<s>`, but it is suffixed by a fragment number. For example, in text A18, the `<s>` formerly numbered 1307, has now been replaced by two `<s>` elements, numbered 1307_1 and 1307_2 respectively.

Fragmentary sentences such as headings or labels in lists are also encoded as `<s>` elements, as in the following example from text CBE:

```
<div type="u">
 <head type="MAIN">
  <s n="835">
   <w c5="AJ0" hw="serious" pos="ADJ">Serious </w>
   <w c5="NN1" hw="fit" pos="SUBST">fit </w>
   <w c5="PRF" hw="of" pos="PREP">of </w>
   <w c5="NN2" hw="giggle" pos="SUBST">giggles</w>
  </s>
 </head>
 <p>
  <s n="836">
   <w c5="AT0" hw="a" pos="ART">A </w>
   <w c5="NN0" hw="pair" pos="SUBST">PAIR </w>
   <w c5="PRF" hw="of" pos="PREP">of </w>
   <w c5="NN1" hw="tv" pos="SUBST">TV </w>
   <w c5="NN2" hw="newsreader" pos="SUBST">newsreaders </w>
     ...</s>...</p>...
</div>
```

As noted above, at the lowest level, the corpus consists of <w> (word) and <c> (punctuation) elements, grouped into <s> (segment) elements. Each <w> element contains three attributes to indicate its morphological class or part of speech, as determined by the CLAWS tagger, a simplified form of that POS code, and an automatically-derived root form or lemma. Each <c> element also carries codes for part of speech, but not for lemma. For example, the word ‘corpora’ wherever it appears in the BNC is presented like this:

```
<w c5="NN2" pos="SUBST" hw="corpus">corpora </w>
```

Any white space following a word in the original source is preserved within the <w> tag, as in the previous example. White space is not added if no space is present in the source, as in the following example:

```
<w c5="NN2" pos="SUBST" hw="corpus">corpora</w>
<c c5="PUN" pos="PUN">.  </c>
```

The <w> element encloses a single token as identified by the CLAWS tagger. Usually this willl correspond with a word as conventionally spelled; there are however two important exceptions. Firstly, CLAWS regards certain common abbreviated or enclitic forms such as ‘’s’ in ‘he’s’ or ‘dog’s’ as distinct tokens, thus enabling it to distinguish them as being an auxiliary verb in the first case, and a genitive marker in the second. For example, ‘It’s’ is encoded as follows:

```
<w c5="PNP" hw="it" pos="PRON">It</w>
<w c5="VBZ" hw="be" pos="VERB">’s </w>
```

while ‘dog’s’ is encoded:

```
<w c5="NN1" hw="dog" pos="SUBST">dog</w>
<w c5="POS" hw="’s" pos="UNK">’s </w>
```

Secondly, CLAWS treats certain common multi-word units as if they were single tokens, giving the whole of a sequence such as ‘in spite of’ a single POS code. These multiword sequences were not distinguished from individual <w> elements in earlier versions of the corpus; in the present version however a new element <mw> (for multiword) has been introduced to mark them explicitly. The individual components of a <mw> sequence are also tagged as <w> elements in the same way as elsewhere. Thus, the phrase ‘in terms of’, which in earlier editions of the BNC would have been encoded as a single <w> element, is now encoded as follows:

```
<mw c5="PRP">
 <w c5="PRP" hw="in" pos="PREP">in </w>
 <w c5="NN2" hw="term" pos="SUBST">terms </w>
 <w c5="PRF" hw="of" pos="PREP">of
 </w>
</mw>
```

Detailed information about the procedures by which the part of speech and lemmatization information was added to the corpus is provided in the full User Reference Guide, available at `http://www.natcorp.ox.ac.uk/XMLedition/URG/posguide.html`. A brief summary of the codes used and their significance is also provided in the reference section below (*7. Word class codes*).

3.5 Editorial indications

Despite the best efforts of its creators, any corpus as large as the BNC will inevitably contain many errors, both in transcription and encoding. Every attempt has been made to reduce the incidence of such errors to an acceptable level, using a number of automatic and semi-automatic validation and correction procedures, but exhaustive proof-reading of a corpus of this size remains economically infeasible. Editorial interventions in the marked up texts take three forms. On a few occasions, where markup or commentary introduced by transcribers during the process of creating the corpus may be helpful to subsequent users, it has been retained in the form of an XML comment. On some occasions, encoders have decided to correct material evidently wrong in their copy text: such corrections are marked using the `<corr>` element. And on several occasions, sampling, anonymization or other concerns, have led to the omission of significant parts of the original source; such omissions are marked by means of the `<gap>` element.

The transcription and editorial policies defined for the corpus may not have been applied uniformly by different transcribers and consequently the usage of these elements is not consistent across all texts. The `<tagsDecl>` element in each text's header may be consulted for an indication of the usage of these and other elements within it (see further section *6.1.6. The encoding description*). Their absence should not be taken to imply that the text is either complete or perfectly transcribed.

In the following example, the first three chapters have been omitted for sampling reasons:

```
<wtext type="FICTION">
 <div level="1" n="1">
  <head>
   <s n="1">
    <w c5="NP0" hw="friday" pos="SUBST">Friday </w>
    <w c5="CRD" hw="16" pos="ADJ">16 </w>
    <w c5="NP0" hw="september" pos="SUBST">September </w>
    <w c5="PRP" hw="to" pos="PREP">to </w>
    <w c5="NP0" hw="tuesday" pos="SUBST">Tuesday </w>
    <w c5="CRD" hw="20" pos="ADJ">20 </w>
    <w c5="NP0" hw="september" pos="SUBST">September</w>
   </s>
  </head>
  <gap desc="chapters 1-3 of book 1" reason="sampling strategy"/>
  <pb n="17"/>
  <div level="2" n="4">
   <p>
    <s n="2">
     <w c5="AV0" hw="once" pos="ADV">Once </w>
     <w c5="AJ0" hw="free" pos="ADJ">free </w>
     <w c5="PRF" hw="of" pos="PREP">of </w>
     <w c5="AT0" hw="the" pos="ART">the </w>
     <w c5="AJ0" hw="knotted" pos="ADJ">knotted </w>
     <w c5="NN2" hw="tentacle" pos="SUBST">tentacles </w>
     <w c5="PRF" hw="of" pos="PREP">of </w>
     <w c5="AT0" hw="the" pos="ART">the
     </w>
     <w c5="AJ0" hw="eastern" pos="ADJ">eastern </w>
     <w c5="NN2" hw="suburb" pos="SUBST">suburbs</w>
<!- ...  ->
    </s>
   </p>...</div>...</div>
</wtext>
<!- C8T ->
```

In the following example, a proper name has been omitted:

```
<s n="547">
 <w c5="PNP" hw="i" pos="PRON">I </w>
 <w c5="VVD" hw="ask" pos="VERB">asked </w>
 <w c5="NP0" hw="mr" pos="SUBST">Mr </w>
 <gap desc="name" reason="anonymization"/>
 <w c5="CJC" hw="and" pos="CONJ">and </w>
...
</s>
<!- KBC ->
```

In the following example, a telephone number has been omitted:

```
<s n="762">
 <w c5="PNP" hw="he" pos="PRON">He </w>
 <w c5="VVD" hw="appeal" pos="VERB">appealed </w>
 <w c5="PRP" hw="for" pos="PREP">for </w>
 <w c5="PNI" hw="anyone" pos="PRON">anyone </w>
 <w c5="PRP" hw="with" pos="PREP">with </w>
 <w c5="NN1" hw="information" pos="SUBST">information </w>
 <w c5="TO0" hw="to" pos="PREP">to </w>
 <w c5="VVI" hw="contact" pos="VERB">contact </w>
 <w c5="PNP" hw="he" pos="PRON">him </w>
 <w c5="AVP" hw="on" pos="ADV">on </w>
 <gap desc="telephone number"/>
 <c c5="PUN">.</c>
</s>
<!- K3C ->
```

In the following example, a typographic error in the original has been corrected:

```
<s n="48">...
<w c5="AJ0" hw="good" pos="ADJ">good </w>
 <w c5="CJC" hw="or" pos="CONJ">or </w>
 <corr sic="herioc">
  <w c5="AJ0" hw="heroic" pos="ADJ">heroic </w>
 </corr>
 <w c5="NN1" hw="behaviour" pos="SUBST">behaviour</w>
...</s>
<!- B17 ->
```

In the following example, a word omitted in the original has been supplied as correction:

```
<s n="1838">
<!- ...  ->
 <w c5="PNP" hw="it" pos="PRON">it </w>
 <w c5="VM0" hw="must" pos="VERB">must </w>
 <corr sic="">
  <w c5="VBI" hw="be" pos="VERB">be
  </w>
 </corr>
 <w c5="VVN" hw="say" pos="VERB">said</w>
```

```
<!- ...  ->
</s>
<!- BNT ->
```

Note that the <sic> element used in some previous editions of the BNC is no longer used.

4 Written texts

4.1 Divisions of written texts

Written texts exhibit a rich variety 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. In the BNC all such structural divisions are represnted uniformly by means of the <div> element, which has attributes n, type, and level.

In written texts, the n attribute is sometimes used to supply an identifying name or number used within the text for a given division, for example, a chapter number, as in the following example:

```
<div type="chapter" n="three" level="1">...</div>
```

More often, however, chapter names or numbers will appear within the text, tagged using the <head> element discussed in section *4.2.1. Headings and captions* below.

The value of the attribute type is used to characterise the function of the textual division (see the reference documentation for the values used). If a value is supplied for one division at a given level, it may be assumed to apply to all subsequent divisions at the same level until the end of the enclosing element, although it is not always explicitly specified.

Where <div> levels are nested, for example where the chapters of a novel are grouped into parts each of which may have its own title or number, the level attribute is used to indicate the depth of nesting. This is not strictly necessary (since an XML-aware processor retains this information) but has been added for the convenience of users of previous versions of the corpus, in which the level was explicitly coded into the name of the surrounding element (<div1>, <div2> etc.)

In text ANY, for example, each chapter of the original novel corresponds with a <div level="2">, because the work contains groups of chapters, each of which begins with a page containing just a date. The opening of the text is therefore encoded as follows:

```
<wtext type="FICTION">
 <div level="1" n="1">
  <s n="1">
   <w c5="NP0" hw="monday" pos="SUBST">Monday</w>
   <c c5="PUN">, </c>
   <w c5="NP0" hw="january" pos="SUBST">January </w>
   <w c5="ORD" hw="13th" pos="ADJ">13th</w>
   <c c5="PUN">, </c>
   <w c5="CRD" hw="1986" pos="ADJ">1986</w>
   <c c5="PUN">.</c>
  </s>
  <div level="2" n="1">
   <p>
    <s n="2">
     <w c5="NP0" hw="victor" pos="SUBST">Victor </w>
     <w c5="NP0" hw="wilcox" pos="SUBST">Wilcox </w>
     <w c5="VVZ" hw="lie" pos="VERB">lies </w>
     <w c5="AJ0" hw="awake" pos="ADJ">awake</w>
         ...</s>
   </p>
  </div>
 </div>...
</wtext>
<!- ANY ->
```

Note however that in some texts initial sentences (like ‘Monday, January 13th, 1986’ above) may have been misplaced, so that they appear at the start of an inner <div> rather than the start of its parent.

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. Note that any prefatory or appended matter not forming part of a text will not generally be captured: the tei elements <front> and <back> elements are not used.

4.2 Paragraph-level elements and chunks

Written texts may be organized into structural units containing more than one <s> element and smaller than any of the divisions discussed in section *4.1. Divisions of written texts* above. The most commonly found such element is the <p> (paragraph), but the following other elements may also appear: <head> (heading), <quote> (quotation), <lg> (line-group or poem), <list> (list), and <bibl> (bibliographic reference).

Each of these elements contains one or more <s> elements, as discussed above; in some cases enclosed by an intermediate element. They are used chiefly to indicate the function of sections of the text, as indicated in the list above.

The following sections provide examples for the use of each of these elements.

4.2.1 Headings and captions

One or more <head> elements of specified types may appear in sequence at the start of any <div> element, or at the start of a <list> or <poem>, as in the following examples:.

```
<div level="1" n="1">
 <head type="MAIN">
  <s n="1">
   <w c5="NN1" hw="ageism" pos="SUBST">AGEISM</w>
  </s>
 </head>
 <head type="SUB">
  <s n="2">
   <w c5="AT0" hw="the" pos="ART">THE </w>
   <w c5="NN1" hw="foundation" pos="SUBST">FOUNDATION </w>
   <w c5="PRF" hw="of" pos="PREP">OF </w>
   <w c5="NN1" hw="age" pos="SUBST">AGE </w>
   <w c5="NN1" hw="discrimination" pos="SUBST">DISCRIMINATION</w>
  </s>
 </head>
 <head type="BYLINE">
  <s n="3">
   <w c5="NP0" hw="steve" pos="SUBST">STEVE </w>
   <w c5="NP0-NN1" hw="scrutton" pos="SUBST">SCRUTTON</w>
  </s>
 </head>...
</div>
<!- B01 ->
```

As shown above, the type attribute is used to distinguish more exactly the function of a heading.

Note that, in the BNC, captions or headings which ‘float’ within the text, that is, which appear elsewhere than at the very beginning of the section which they name, are not encoded as <head> elements. A <head> element can appear only at the start of a text division and is logically associated with it (for example, chapter titles, newspaper headlines etc.). Paragraphs which provide heading or captioning information, but which are logically independent of their position within a textual division (for example, captions attached to pictures or figures, or ‘pull-quotes’ embedded within the text) are represented in the same way as any other paragraph of text, using the <p> element, but specifying the value `caption` in their type attribute.

In the following example, a paragraph of captioning information interrupts the narrative:

```
<p>...
<s n="151">
  <w c5="DT0" hw="some" pos="ADJ">Some </w>
  <w c5="AJ0" hw="related" pos="ADJ">related </w>
  <w c5="NN2" hw="form" pos="SUBST">forms </w>
  <w c5="VBB" hw="be" pos="VERB">are </w>
  <w c5="AV0" hw="well" pos="ADV">better </w>
  <w c5="VVN" hw="study" pos="VERB">studied </w>
  <w c5="PRP" hw="in" pos="PREP">in </w>
  <w c5="AJ0" hw="polished" pos="ADJ">polished </w>
  <w c5="NN1" hw="section" pos="SUBST">section </w>
  <w c5="PRP" hw="like" pos="PREP">like </w>
  <w c5="AT0" hw="the" pos="ART">the </w>
  <w c5="NN2" hw="coral" pos="SUBST">corals</w>
  <c c5="PUN">.</c>
 </s>
</p>
<p type="caption">
 <s n="152">
  <w c5="NN1-VVB" hw="section" pos="SUBST">Section </w>
  <w c5="PRP" hw="across" pos="PREP">across </w>
  <w c5="AT0" hw="a" pos="ART">a </w>
  <w c5="AJ0" hw="typical" pos="ADJ">typical </w>
  <w c5="NN1" hw="rugose" pos="SUBST">rugose </w>
  <w c5="NN1" hw="coral" pos="SUBST">coral </w>
  <c c5="PUL">(</c>
  <hi rend="it">
   <w c5="NP0" hw="palaeosmilia" pos="SUBST">Palaeosmilia </w>
  </hi>
  <c c5="PUN">, </c>
  <w c5="AJ0" hw="carboniferous" pos="ADJ">Carboniferous</w>
  <c c5="PUR">)</c>
 </s>
</p>
<p>
 <s n="153">
  <w c5="AT0" hw="a" pos="ART">A </w>
  <w c5="NN1" hw="coral" pos="SUBST">coral </w>
  <w c5="VBZ" hw="be" pos="VERB">is </w>
  <w c5="AV0" hw="essentially" pos="ADV">essentially </w>
  <w c5="AT0" hw="a" pos="ART">a </w>
  <w c5="NN1" hw="sea" pos="SUBST">sea </w>
  <w c5="NN1" hw="anemone" pos="SUBST">anemone </w>
  <w c5="DTQ" hw="which" pos="PRON">which </w>
  <w c5="VVZ" hw="support" pos="VERB">supports </w>
  <w c5="DPS" hw="it" pos="PRON">its </w>
  <w c5="NN1" hw="body" pos="SUBST">body </w>
  <w c5="PRP" hw="by" pos="PREP">by </w>
  <w c5="AT0" hw="a" pos="ART">a </w>
  <w c5="NN1" hw="skeleton" pos="SUBST">skeleton </w>
  <w c5="PRF" hw="of" pos="PREP">of </w>
  <w c5="NN1" hw="calcium" pos="SUBST">calcium </w>
  <w c5="NN1" hw="carbonate" pos="SUBST">carbonate</w>
  <c c5="PUN">.</c>
 </s>
...
</p>
<!- AMM->
```

4.2.2 Quotations

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 is marked up using the <quote> element. This may contain any combination of other chunks (for example paragraphs, poems, lists) but may not directly contain <w> or <s> elements. A reference for the citation may also be contained within it, usually marked by means of the <bibl> element, as in the following example:

```
<quote>
 <p>
  <s n="2">
   <w c5="AV0" hw="now" pos="ADV">Now </w>
   <w c5="NN1" hw="hatred" pos="SUBST">hatred </w>
   <w c5="VBZ" hw="be" pos="VERB">is </w>
   <mw c5="AV0">
    <w c5="PRP" hw="by" pos="PREP">by </w>
    <w c5="AV0" hw="far" pos="ADV">far </w>
   </mw>
   <w c5="AT0" hw="the" pos="ART">the </w>
   <w c5="AJS" hw="long" pos="ADJ">longest </w>
   <w c5="NN1" hw="pleasure" pos="SUBST">pleasure</w>
   <c c5="PUN">; </c>
   <w c5="NN2" hw="man" pos="SUBST">Men </w>
   <w c5="VVB-NN1" hw="love" pos="VERB">love </w>
   <w c5="PRP-AVP" hw="in" pos="PREP">in </w>
   <w c5="NN1" hw="haste" pos="SUBST">haste </w>
   <w c5="CJC" hw="but" pos="CONJ">but </w>
   <w c5="PNP" hw="they" pos="PRON">they </w>
   <w c5="VVB" hw="detest" pos="VERB">detest </w>
   <w c5="PRP" hw="at" pos="PREP">at </w>
   <w c5="NN1" hw="leisure" pos="SUBST">leisure</w>
   <c c5="PUN">...</c>
  </s>
  <bibl rend="it">
   <s n="3">
    <w c5="NP0" hw="lord" pos="SUBST">Lord </w>
    <w c5="NP0" hw="byron" pos="SUBST">Byron</w>
   </s>
  </bibl>
 </p>
</quote>
<!- G01 ->
```

4.2.3 Poetry

Poetry is distinguished from prose in the BNC where it is so presented in the original, for example as fragments of verse or song appearing within or between paragraphs of prose. The <l> (line) element is used to mark each verse line; where there are several such lines, perhaps with a heading, they are grouped together using the <lg> (linegroup) element, and any title or heading present is marked with a <head> element.

For example:

```
<p>
 <s n="472">
  <w c5="AV0" hw="thus" pos="ADV">Thus </w>
  <w c5="CJS" hw="when" pos="CONJ">when </w>
  <w c5="NP0" hw="burns" pos="SUBST">Burns </w>
  <w c5="VVD" hw="write" pos="VERB">wrote</w>
  <c c5="PUN">:</c>
 </s>
 <lg>
  <l>
   <s n="473">
    <w c5="AT0" hw="the" pos="ART">The </w>
    <w c5="NN1" hw="rank" pos="SUBST">rank </w>
    <w c5="VBZ" hw="be" pos="VERB">is
    </w>
    <w c5="CJC" hw="but" pos="CONJ">but </w>
    <w c5="AT0" hw="the" pos="ART">the </w>
    <w c5="NN1" hw="guinea" pos="SUBST">guinea </w>
    <w c5="NN1-VVB" hw="stamp" pos="SUBST">stamp</w>
    <c c5="PUN">, </c>
    <w c5="AT0" hw="the" pos="ART">The </w>
    <w c5="NN1" hw="man" pos="SUBST">man</w>
    <w c5="VBZ" hw="be" pos="VERB">'s </w>
    <w c5="AT0" hw="the" pos="ART">the </w>
    <w c5="NN1" hw="gold" pos="SUBST">gold
    </w>
    <w c5="PRP" hw="for" pos="PREP">for </w>
    <w c5="NN1-AJ0" hw="a'" pos="SUBST">a' </w>
    <w c5="DT0" hw="that" pos="ADJ">that</w>
   </s>
  </l>
 </lg>
</p>
<!- ECV ->
```

Note that the <l> element is not used to mark typographic lineation. Layout information is not, in general, preserved in the BNC.

4.2.4 Lists

A list is a collection of distinct items flagged as such by special layout in written texts, often functioning as a single syntactic unit. Lists may appear within or between paragraphs. Where marked, lists are tagged with the <list> element, which may contain the following subelements:

A <list> element consists of an optional <head> element, followed by one or more <item> elements, each of which may optionally be preceded by a <label> element, 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.

The <item> element may appear only inside lists. It contains the same mixture of elements as a paragraph, and may thus contain one or more nested lists. It may also contains a series of paragraphs, each marked with a <p> element.

Here is an example of a simple list:

```
<s n="1716">
 <w c5="AT0" hw="the" pos="ART">The </w>
 <w c5="NN2" hw="personnel" pos="SUBST">personnel </w>
 <w c5="NN1-VVB" hw="file" pos="SUBST">file </w>
 <w c5="VM0" hw="might" pos="VERB">might </w>
 <w c5="VBI" hw="be" pos="VERB">be </w>
 <w c5="VVN" hw="examine" pos="VERB">examined </w>
 <w c5="TO0" hw="to" pos="PREP">to </w>
 <w c5="VVI" hw="list" pos="VERB">list </w>
 <w c5="DT0" hw="all" pos="ADJ">all </w>
 <w c5="AT0" hw="the" pos="ART">the </w>
 <w c5="NN2" hw="employee" pos="SUBST">employees </w>
 <w c5="PNQ" hw="who" pos="PRON">who </w>
 <w c5="VVB" hw="meet" pos="VERB">meet </w>
 <w c5="AT0" hw="the" pos="ART">the </w>
 <w c5="AJ0" hw="following" pos="ADJ">following </w>
 <w c5="NN2" hw="criterion" pos="SUBST">criteria</w>
 <c c5="PUN">:</c>
</s>
<list>
 <item>
  <s n="1717">
   <w c5="VVZ" hw="speak" pos="VERB">Speaks </w>
   <w c5="NN1" hw="japanese" pos="SUBST">Japanese </w>
   <w c5="CJC" hw="and" pos="CONJ">AND</w>
  </s>
 </item>
 <item>
  <s n="1718">
   <w c5="VVB-NN1" hw="graduate" pos="VERB">Graduate </w>
   <w c5="NN1" hw="engineer" pos="SUBST">Engineer </w>
   <w c5="CJC" hw="and" pos="CONJ">AND</w>
  </s>
 </item>
 <item>
  <s n="1719">
   <w c5="AJ0-NN1" hw="single" pos="ADJ">Single</w>
  </s>
 </item>
</list>
<!- FPG ->
```

Here is an example of a labelled list:

```
<s n="24">
 <w c5="DT0" hw="this" pos="ADJ">This </w>
 <w c5="NN1" hw="expression" pos="SUBST">expression </w>
 <w c5="VBZ" hw="be" pos="VERB">is </w>
 <w c5="VVN" hw="derive" pos="VERB">derived </w>
 <w c5="VVG" hw="assume" pos="VERB">assuming</w>
</s>
<list>
 <label>
  <s n="25">
   <c c5="PUL">(</c>
   <w c5="ZZ0" hw="a" pos="SUBST">a</w>
   <c c5="PUR">)</c>
  </s>
 </label>
 <item>
  <s n="26">
   <w c5="AT0" hw="the" pos="ART">the </w>
   <w c5="NN1" hw="volume" pos="SUBST">volume </w>
   <w c5="NN1-VVB" hw="change" pos="SUBST">change </w>
   <w c5="PRP" hw="on" pos="PREP">on </w>
   <w c5="NN1-VVG" hw="mixing" pos="SUBST">mixing </w>
   <gap desc="formula"/>
   <c c5="PUN">,</c>
  </s>
 </item>
 <label>
  <s n="27">
   <c c5="PUL">(</c>
   <w c5="ZZ0" hw="b" pos="SUBST">b</w>
   <c c5="PUR">)</c>
  </s>
 </label>
 <item>
  <s n="28">
   <w c5="AT0" hw="the" pos="ART">the </w>
   <w c5="NN2" hw="molecule" pos="SUBST">molecules </w>
   <w c5="VBB" hw="be" pos="VERB">are </w>
   <w c5="DT0" hw="all" pos="ADJ">all </w>
   <w c5="PRF" hw="of" pos="PREP">of </w>
   <w c5="AJ0" hw="equal" pos="ADJ">equal </w>
   <w c5="NN1" hw="size" pos="SUBST">size</w>
   <c c5="PUN">,</c>
  </s>
 </item>
 <label>
  <s n="29">
   <c c5="PUL">(</c>
   <w c5="ZZ0" hw="c" pos="SUBST">c</w>
   <c c5="PUR">)</c>
  </s>
 </label>
 <item>
  <s n="30">
   <w c5="DT0" hw="all" pos="ADJ">all </w>
   <w c5="AJ0" hw="possible" pos="ADJ">possible </w>
   <w c5="NN2" hw="arrangement" pos="SUBST">arrangements </w>
   <w c5="VHB" hw="have" pos="VERB">have </w>
   <w c5="AT0" hw="the" pos="ART">the </w>
   <w c5="DT0" hw="same" pos="ADJ">same </w>
   <w c5="NN1" hw="energy" pos="SUBST">energy</w>
   <c c5="PUN">,</c>
   <c c5="PUN">, </c>
   <w c5="CJC" hw="and" pos="CONJ">and</w>
  </s>
 </item>
 <label>
  <s n="31">
   <c c5="PUL">(</c>
   <w c5="ZZ0" hw="d" pos="SUBST">d</w>
   <c c5="PUR">)</c>
  </s>
 </label>
 <item>
  <s n="32">
   <w c5="AT0" hw="the" pos="ART">the </w>
```

```
   <w c5="NN1" hw="motion" pos="SUBST">motion </w>
   <w c5="PRF" hw="of" pos="PREP">of </w>
   <w c5="AT0" hw="the" pos="ART">the </w>
   <w c5="NN2" hw="component" pos="SUBST">components </w>
   <w c5="PRP" hw="about" pos="PREP">about </w>
   <w c5="DPS" hw="they" pos="PRON">their </w>
   <w c5="NN1" hw="equilibrium" pos="SUBST">equilibrium </w>
   <w c5="NN2" hw="position" pos="SUBST">positions </w>
   <w c5="VVZ" hw="remain" pos="VERB">remains </w>
   <w c5="AJ0" hw="unchanged" pos="ADJ">unchanged </w>
   <w c5="PRP" hw="on" pos="PREP">on </w>
   <w c5="NN1-VVG" hw="mixing" pos="SUBST">mixing</w>
  </s>
 </item>
</list>
<!- HRG->
```

4.2.5 Bibliographic references

The <bibl> element is chiefly used in the header of each corpus file, to document its source. Bibliographic citations or references associated with a quotation are however also marked in the same way in a few of the corpus texts, as shown above, or as in the following example:

```
<quote>
 <p>
  <s n="426">
   <w c5="CJC" hw="and" pos="CONJ">And </w>
   <w c5="DT0" hw="all" pos="ADJ">all </w>
   <w c5="AT0" hw="the" pos="ART">the </w>
   <w c5="NN2" hw="undertaking" pos="SUBST">undertakings </w>
   <w c5="PNP" hw="we" pos="PRON">we </w>
   <w c5="VM0" hw="might" pos="VERB">might </w>
   <w c5="VVI" hw="speak" pos="VERB">speak </w>
   <w c5="PRF" hw="of" pos="PREP">of </w>
   <w c5="VVB" hw="reduce" pos="VERB">reduce </w>
   <w c5="PNX" hw="themselves" pos="PRON">themselves </w>
   <w c5="PRP" hw="to" pos="PREP">to </w>
   <w c5="AT0" hw="a" pos="ART">a </w>
   <w c5="AJ0" hw="single" pos="ADJ">single </w>
   <w c5="PNI" hw="one" pos="PRON">one</w>
   <c c5="PUN">, </c>
   <w c5="DT0" hw="that" pos="ADJ">that </w>
   <w c5="PRF" hw="of" pos="PREP">of </w>
   <hi rend="it">
    <w c5="VVG" hw="make" pos="VERB">making </w>
    <w c5="NN1" hw="history" pos="SUBST">history </w>
   </hi>
   <c c5="PUN">.</c>
  </s>
  <bibl>
   <s n="427">
    <w c5="NP0" hw="jean-paul" pos="SUBST">Jean-Paul </w>
    <w c5="NP0" hw="sartre" pos="SUBST">Sartre</w>
   </s>
  </bibl>
 </p>
</quote>
```

Note that the <bibl> element used within corpus texts has none of the more detailed sub-elements described for it in *6.1.5.2. Structured bibliographic record.* Like all the other elements described in the present subsection, the <bibl> element appearing within corpus texts contains only <s> elements.

4.3 Phrase-level elements

Phrase-level elements are elements which cannot appear directly within a textual division, but must be contained by some other element. In practice, this means they will be contained within an <s> element. In addition to the <w>, <mw>, and <c> elements already discussed, there are two phrase-level elements which may appear within <s> elements in written texts: <pb> (for page breaks)and <hi> (for highlighted text).

4.3.1 Page breaks

Wherever possible, the original pagination and page numbering of the source text has been preserved. The <pb> element is used to mark the approximate position in the text at which each new page starts, and its n attribute supplies the number of the page.

```
<l>
 <s n="1403">
  <c c5="PUN">-- </c>
  <w c5="CJC" hw="and" pos="CONJ">and </w>
  <w c5="NN2" hw="creditor" pos="SUBST">creditors </w>
  <w c5="VVB" hw="grow" pos="VERB">grow </w>
  <w c5="AJ0" hw="cruel" pos="ADJ">cruel</w>
  <c c5="PUN">,</c>
 </s>
</l>
<l>
 <s n="1404">
  <pb n="75"/>
  <w c5="AV0" hw="so" pos="ADV">so </w>
  <w c5="PNP" hw="he" pos="PRON">he </w>
  <w c5="VVZ" hw="bow" pos="VERB">bows </w>
  <w c5="CJC" hw="and" pos="CONJ">and </w>
  <w c5="NN2-VVZ" hw="scrape" pos="SUBST">scrapes</w>
  <c c5="PUN">,</c>
 </s>
</l>
<!- HNU ->
```

Where several pages have been left out of a transcription, for example because they are blank or contain illustrations only, a <pb> element may be given for each, as in this example:

```
<s n="1323">
 <w c5="PNP" hw="i" pos="PRON">I </w>
 <w c5="VHB" hw="have" pos="VERB">have</w>
 <w c5="XX0" hw="not" pos="ADV">n’t </w>
 <w c5="VBN" hw="be" pos="VERB">been </w>
 <w c5="PRP" hw="to" pos="PREP">to </w>
 <w c5="AT0" hw="an" pos="ART">an </w>
 <w c5="AJ0" hw="organized" pos="ADJ">organized </w>
 <w c5="NN1" hw="campsite" pos="SUBST">campsite </w>
 <w c5="PRP" hw="for" pos="PREP">for </w>
 <pb n="64"/>
 <pb n="65"/>
 <pb n="66"/>
 <w c5="AV0" hw="perhaps" pos="ADV">perhaps </w>
 <w c5="CRD" hw="fifteen" pos="ADJ">fifteen </w>
 <w c5="NN2" hw="year" pos="SUBST">years</w>
 <c c5="PUN">, </c>
 <w c5="AV0" hw="so" pos="ADV">so </w>
 <w c5="DT0" hw="all" pos="ADJ">all </w>
 <w c5="DT0" hw="this" pos="ADJ">this </w>
 <w c5="VBZ" hw="be" pos="VERB">is </w>
 <w c5="AJ0" hw="new" pos="ADJ">new </w>
 <w c5="PRP" hw="to" pos="PREP">to </w>
 <w c5="PNP" hw="i" pos="PRON">me</w>
 <c c5="PUN">.</c>
</s>
<!- A6T ->
```

4.3.2 Highlighted phrases

Typographic changes or highlighting in the original may not be marked in the transcript at all. Alternatively, highlighted phrases, and the kind of highlighting used, may be recorded in one of two ways:

- using the rend (rendition) attribute on elements for which this is defined
- using the <hi> (highlighted) element

The former is used where the whole of the content of one of the elements <bibl>, <corr>, <div>, <head>, <item>, <l>, <label>, <list>, <p>, <quote> or <stage> is highlighted. The latter is used on all other occasions. The values used for the rend attribute in BNC Baby are as follows:

Table 8: Rendition values used in BNC Baby

code	meaning
bo	bold face
it	italic font
hi	superscript
lo	subscript
ul	underlined

The <hi> element is frequently used for text that is italicized in the original source because it is a title, a foreign word or phrase, a technical term, etc. The current version of the markup does not however distinguish amongst these functions, as shown in the following examples:

```
<s n="36">He was always searching, rummaging:
<hi rend="it">Les vrais paradis sont les paradis
   qu’on a perdus</hi>.</s>
<!- FAJ ->
```

```
<s n="295">"Oh, so there will be
<hi rend="it">two </hi>new volumes,
will there?"</s>
<!-H9D ->
```

```
<p>
 <s n="167">Wivenhoe gemmologist Stephanie
   Coward,<hi rend="it">above </hi>, is showing
   her work at the Pam Schomberg Gallery, St John’s Street, Colchester.</s>
<!- CFC ->
</p>
```

It should be noted that the purpose of the rend attribute is *not* to provide information adequate to the needs of a typesetter, but simply to record some qualitative information about the original.

Like all other phrase-level elements, each <hi> element must be entirely contained by an <s> element. This implies that where, for example, a bolded passage contains more than one sentence, 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.

5 Spoken texts

5.1 Basic structure: spoken texts

The spoken material transcribed for the BNC is also organized into ‘texts’, which are subdivided into ‘divisions’, made up of <w> and <mw> elements grouped into <s> elements in the same way as written texts. However there a number of other elements specific to spoken texts, and their hierarchic organization is naturally not the same as that of written texts. For this reason, a different element (<stext>) is used to represent a spoken text.

In demographically sampled spoken texts, each distinct conversation recorded by a given respondent is treated as a distinct <div> element. All the conversations from a single respondent are then grouped together to form a single <stext> element. Context-governed spoken texts, however, do not use the <div> element: the <stext> element for a context governed text is composed only of <u> elements, not grouped into any unit smaller than the <stext> itself.

The <s> elements making up a spoken text are grouped not into <p> or other similar elements, but instead into <u> elements. Each <u> (utterance) element marks a stretch of uninterrupted speech from a given speaker; (see section *5.2. Utterances*). Interspersed within and between <u> elements, a variety of other elements indicate para-linguistic phenomena noticed by the transcribers (see section *5.3. Paralinguistic phenomena*).

The methods and principles applied in transcription and normalisation of speech are defined in a BNC working paper TGCW21 *Spoken Corpus Transcription Guide*, and have also been described in subsequent publications (e.g. Crowdy 1994). The editorial tags discussed in section *3.5. Editorial indications* above are also used to represent normalisation practice when dealing with transcribed speech.

5.2 Utterances

The term utterance is used in the BNC to refer to a continuous stretch of speech produced by one participant in a conversation, or by a group of participants. Structurally, the corresponding element behaves in a similar way to the <p> element in a written text — it groups a sequence of <s> elements together.

The who attribute is required on every <u>: its function is to identify the person or group of people making the utterance, using the unique code defined for that person in the appropriate section of the header. A simple example follows:

```
<u who="PS0H7">
 <s n="3778">
  <w c5="ITJ" hw="mm" pos="INTERJ">Mmm </w>
  <w c5="ITJ" hw="mm" pos="INTERJ">mm</w>
  <c c5="PUN">.</c>
 </s>
</u>
<!- KCV ->
```

The code PS1LW used here will be specified as the value for the xml:id attribute of some <person> element within the header of the text from which this example is taken. The code PS000 is used where the speaker cannot be identified and the code `PS001` is used for a group of unidentified speakers. Where there are several distinct, but unidentified, speakers within a text, distinct identifiers are used. For example, if text `XYZ` contains two different but unidentified speakers, one of them will be given the identifier `XYZSP001`, and the other `XYZSP002`.

5.3 Paralinguistic phenomena

In transcribing spoken language, it is necessary to select from the possibly very large set of distinct paralinguistic phenomena which might be of interest. In the texts transcribed for the BNC, encoders were instructed to mark the following such phenomena:

voice quality for example, whispering, laughing, etc., both as discrete events and as changes in voice quality affecting passages within an utterance.

non-verbal but vocalised sounds for example, coughs, humming noises etc.

non-verbal and non-vocal events for example passing lorries, animal noises, and other matters considered worthy of note.

significant pauses silence, within or between utterances, longer than was judged normal for the speaker or speakers.

unclear passages whole utterances or passages within them which were inaudible or incomprehensible for a variety of reasons.

speech management phenomena for example truncation, false starts, and correction.

overlap points at which more than one speaker was active.

Other aspects of spoken texts are not explicitly recorded in the encoding, although their headers contain considerable amounts of situational and participant information.

In many cases, because no standardized set of descriptions was predefined, transcribers gave very widely differing accounts of the same phenomena. An attempt has however been made to normalize the descriptions for some of these elements in the BNC XML editions.

The following elements are used to mark these phenomena:

event a non-verbal and non-vocal event, such as a door slamming, occurring during conversation and regarded as significant by the transcriber

pause a marked pause during or between utterances

shift a marked change in voice quality for any one speaker

trunc a word or phrase truncated during speech

unclear a point in a spoken text at which the transcriber has been unable to determine what is happening, e.g. who is speaking or what is being said

vocal a non-linguistic but communicative sound made by a speaker, such as a sharp intake of breath or cough.

With the exception of `<trunc>` and `<shift>`, these elements may all carry a dur attribute to indicate the duration of the phenomenon concerned. Its value is normally specified only if it is greater than 5 seconds, and its accuracy is only approximate.

With the exception of the `<trunc>` element, which is a special case of the editorial tags discussed in section *3.5. Editorial indications* above, all of these elements are empty, and may appear anywhere within a transcription.

The following example shows how the presence of the `<event>` tag can sometimes help make sense of otherwise seemingly random bits of conversation:

```
<u who="PS1A9">
 <s n="775">What are you doing?</s>
 <s n="776">
  <event desc="dog barks"/>My giddy aunt!</s>
 <s n="777">Are you playing rugby this afternoon Kevin?</s>
</u>
<!- KBC ->
```

(Note that in this and subsequent examples all word-level tagging has been suppressed in order to aid legibility)

The values used for the desc attribute of the `<event>` and `<vocal>` elements are free text strings. Typical values frequently found in BNC Baby are listed below:

As noted above, a distinction is made between discrete vocal events, such as laughter, and changes in voice quality, such as words which are spoken in a laughing tone. The former are encoded using the `<vocal>` element, as in the following example:

```
<u who="PS0Y5"><s n="49"><vocal desc="laugh"/>Right <unclear/>.</s>
</u>
<!- KB5 ->
```

The `<shift>` element is used instead where the laughter indicates a change in voice quality, as in the following example:

```
<u who="PS02G"><s n="10649"><shift new="laughing"/>Good
job I didn't have to read it <shift/>!</s></u>
<!- KB7 ->
```

Here the passage between the tags `<shift new="laughing">` and `<shift>` is spoken with a laughing intonation.

A list of the most frequent values used for the new attribute in BNC Baby is given below; these eight values account for 3162 of the 3295 occurrences (95.6%).

Table 9: Voice quality codes

code	count	%
laughing	1522	46.2
reading	606	18.4
singing	449	13.6
shouting	276	8.37
whispering	129	3.91
yawning	71	2.15
crying	62	1.89
screaming	47	1.42

The `<trunc>` element is used to enclose fragmentary words caused by repair or hesitation, as in the following example:

```
<u who="PS03W"><s n="757">Then <trunc>Mar </trunc>this
guy called Mark is there I think.</s>
<!- KBD ->
```

5.4 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 `<align>` element is used to mark synchronization points.It carries a with attribute, the value of which may be thought of as identifying some point in time. Where two or more `<align>` elements specify the same value for this attribute, their locations are assumed to be synchronised.

The following example demonstrates how this mechanism is used to indicate that the second speaker (PS02G) speaks simultaneously with the first, starting from the words "and it's not far from town":

```
<u who="PS02H">
 <s n="126">
  <w c5="AJ0" hw="handy" pos="ADJ">Handy</w>
  <c c5="PUN">, </c>
  <w c5="AV0" hw="really" pos="ADV">really</w>
  <c c5="PUN">.</c>
 </s>
 <s n="127">
  <pause/>
  <align with="KB7LC00S"/>
  <w c5="CJC" hw="and" pos="CONJ">And </w>
  <w c5="PNP" hw="it" pos="PRON">it</w>
  <w c5="VBZ" hw="be" pos="VERB">'s </w>
  <w c5="XX0" hw="not" pos="ADV">not </w>
  <w c5="AJ0" hw="far" pos="ADJ">far </w>
  <w c5="PRP" hw="from" pos="PREP">from </w>
  <w c5="NN1" hw="town" pos="SUBST">town</w>
  <c c5="PUN">.</c>
 </s>
</u>
<u who="PS02G">
 <s n="128">
  <align with="KB7LC00S"/>
  <w c5="AV0" hw="very" pos="ADV">Very </w>
  <w c5="AJ0" hw="handy" pos="ADJ">handy</w>
  <c c5="PUN">.</c>
 </s>
 <s n="129">
  <w c5="PNP" hw="you" pos="PRON">You </w>
  <w c5="VM0" hw="can" pos="VERB">can </w>
  <w c5="VVI" hw="go" pos="VERB">go </w>
  <w c5="AVP" hw="in" pos="ADV">in </w>
  <w c5="PRP" hw="for" pos="PREP">for </w>
  <w c5="AT0" hw="a" pos="ART">a </w>
  <w c5="NN1" hw="drink" pos="SUBST">drink</w>
  <c c5="PUN">.</c>
 </s>
</u>
<u who="PS02H"/>
<!- KB7 ->
```

6 The header

The header of a TEI-conformant text provides a structured description of its contents, analogous to the title page and front matter of a book. The component elements of a TEI header are intended to provide in machine-processable form all the information needed to make sensible use of the Corpus.

Every separate text in BNC Baby (i.e. each <bncDoc> element) has its own header, referred to below as a text header. In addition, the corpus itself has a header, referred to below as the corpus header, containing information which is applicable to the whole corpus. Both corpus and text headers are represented by <teiHeader> elements.

The corpus header is supplied in a separate file called bncHdr.xml, whereas text headers are prefixed to each file in the Texts directory. A TEI header contains a file description (section *6.1. The file description*), an encoding description (section *6.1.6. The encoding description*), a profile description (section *6.1.7. The profile description*) and a revision description (section *6.1.8. The revision description*).

6.1 The file description

The file description (<fileDesc>) is the first of the four main constituents of the header. It is intended to document an electronic file i.e. (in the case of a corpus header) the whole corpus, or (in the case of a text header) any characteristics peculiar to an individual file within it. In each case, it contains a title statement (<titleStmt>), followed by an edition statement (<editionStmt>), and extent statement (<extent>), a publication statement (<publicationStmt>) and a source description (<sourceDesc>), in that order.

Further detail for each of these is given in the following subsections.

6.1.1 The title statement

The title statement (<titleStmt>) element of a BNC text contains one or more <title> elements, optionally followed by <author>, <editor>, or <respStmt> elements. These sub-elements are used throughout the header, wherever the title of a work or a statement of responsibility are required.

For the BNC Baby corpus header, the title statement looks like this:

```
<titleStmt>
 <title>BNC Baby:  a sampling of the British National Corpus</title>
 <respStmt>
  <resp>Selection, design, distribution</resp>
  <name>Research Technologies Service, Oxford University</name>
 </respStmt>
 <respStmt>
  <resp>Creation of original BNC
  </resp>
  <name>BNC Consortium (OUP, Longman, UCREL, OUCS, Chambers)
  </name>
 </respStmt>
</titleStmt>
```

In individual corpus texts, the title statement follows a pattern like the following:

```
<titleStmt>
 <title>The National Trust Magazine.  Sample containing
   about 21015 words from a periodical (domain:  arts)</title>
 <respStmt>
  <resp> Data capture and transcription </resp>
  <name>Oxford University Press </name>
 </respStmt>
</titleStmt>
```

The content of the <title> element includes the title of the source, followed by the phrase "Sample containing about", the approximate word count for the sample, and further information about the text type and domain, all extracted from other parts of the header. This is followed by responsibility statements showing which of the BNC Consortium members was responsible for capturing the text originally.

Here are some typical examples:

```
<title> Man at the sharp end.  Sample containing
about 37622 words from a book (domain:  imaginative)</title>
```

```
<title> Belfast Telegraph:  Religious affairs stories.  Sample
containing about 2180 words from a periodical (domain:  belief and
thought)</title>
```

```
<title> 6 conversations recorded by 'Pauline' (PS0N3) between
21 and 24 February 1992 with 8 interlocutors, totalling 1668 s-units, 16234
words, and over 1 hour 49 minutes 24 seconds of recordings.</title>
```

A <respStmt> element is used to indicate each agency responsible for any significant effort in the creation of the text. Since responsibilities for data encoding and storage, and for enrichment, are the same for all texts, they are not repeated in each text header. The responsibility for original data capture and transcription varies text by text, and is therefore stated in each text header, in the following manner:

```
<respStmt>
 <resp>Data capture and transcription
 </resp>
 <name>Longman ELT
 </name>
</respStmt>
```

Author and editor information for the source from which a text is derived (e.g. the author of a book) is not included in the <filedesc> element but in the <sourceDesc> element discussed below (*6.1.5. The source description*).

6.1.2 The edition statement

The <editionStmt> element is used to specify an edition for each file making up the corpus. For the corpus header, this takes the form: Second Edition since this the second edition of BNC Baby. For individual texts, since the header has been taken inchanged from the BNC XML edition, in each file it takes the form:

```
<editionStmt>
 <edition>BNC XML Edition, January 2007</edition>
</editionStmt>
```

6.1.3 The extent statement

The <extent> element is used in each text header to specify the size of the text to which it is attached, as in the following example:

```
<extent> 21015 tokens; 21247 w-units; 957 s-units </extent>
```

These counts do not include the size of the header itself. The number of 'tokens' is generated by the Unix wc utility, which simply counts blank delimited strings; the other figures give the number of <w> and <s> elements respectively.

6.1.4 The publication statement

The <publicationStmt> element is used to specify publication and availability information for an electronic text. It contains information about the name and address of the distributor, identification numbers etc., notes on availability and publication dates.

Individual text headers contains the following fixed text for the first two of these:

```
<distributor>Distributed under licence by Oxford University Computing
Services on behalf of the BNC Consortium.</distributor>
<availability> This material is protected by international copyright
laws and may not be copied or redistributed in any way.  Consult the BNC
Web Site at http://www.natcorp.ox.ac.uk for full licencing and distribution
conditions.</availability>
```

For contractual reasons, the corpus header includes a somewhat longer rehearsal of the terms and conditions under which the BNC is made available.

For individual text headers, two identification numbers are supplied, distinguished by the value of their type attribute.

```
<idno type="bnc">A0A</idno>
<idno type="old">CAMfct</idno>
```

The second identifier (of type old) is the old-style mnemonic or numeric code attached to BNC texts during the production of the corpus, and is still used to label the original printed source materials in the BNC Archive. The first three character code (of type bnc) is the standard BNC identifier. It is also used both for the filename in which the text is stored and as the value supplied for the xml:id attribute on the <bncDoc> element containing the whole text, and should always be used to cite the text. The code is a completely arbitrary identifier, and does not indicate anything about the nature of the text.

6.1.5 The source description

The <sourceDesc> element is used to supply bibliographic details for the original source material from which a corpus text derives in individual text headers. In the case of a BNC text, this might be a book, pamphlet, newspaper etc., or a recording. In the latter case, a <recordingStmt> is used; otherwise, the <bibl> element is used.

All bibliographic data supplied in the individual text headers is collected together and reproduced in section «bibliog» below.

The recording statement

The recording statement (<recordingStmt>) element contains one or more <recording> elements, one for each tape transcribed.

In the following simple example, typical of most of the ‘context-governed’ parts of the BNC, the <recording> element has no content at all:

```
<recordingStmt>
 <recording
   n="121101"
   date="1994-02-09"
   time="11:00"
   type="DAT"/>
</recordingStmt>
```

When, as is often the case for the spoken demographic parts of the BNC, a text has been made up by transcribing several different recordings made by a single respondent over a period of time, each such recording will have its own <recording> element, as in the following example:

```
<recordingStmt>
 <recording
   n="018201"
   dur="322"
   date="1991-11-28"
   time="18:15+"
   type="Walkman"
   xml:id="KB7RE000"/>
 <recording
   n="018202"
   dur="253"
   date="1991-11-28"
   time="18:15+"
   type="Walkman"
   xml:id="KB7RE001"/>
<!- ...  ->
 <recording
   n="018207"
   dur="630"
   date="1991-11-29"
   time="10:15+"
   type="Walkman"
   xml:id="KB7RE006"/>
 <recording
   n="018301"
   dur="75"
   date="1991-11-29"
   time="12:15+"
   type="Walkman"
   xml:id="KB7RE007"/>
<!- ...  ->
</recordingStmt>
```

Note the presence of an xml:id attribute on each of the above recordings. The value given here is used to indicate the recording from which a given part of the text was transcribed. Each recording is transcribed as a distinct `<div>` (division) element within an `<stext>`. In that element, the identifier of the source recording is supplied as the value of a decls attribute. Thus, in the spoken text derived from the above mentioned recordings, there will be a `<div>` element starting as follows:

```
<div decls="KB7RE0077"> ...</div>
```

which will contain the part of text transcribed from that recording. The identifier supplied on the n attribute is quite distinct; it may be used to identify the original tape recording was made within the British Library's Sound Archive.

Structured bibliographic record

In addition to its usage within the corpus texts (see *4.2.5. Bibliographic references*), the `<bibl>` element is also used to record bibliographic information for each non-spoken component of the BNC. In this case, its structure is constrained to contain only the following elements in the order specified: `<title>`, `<editor>`, `<author>`, `<imprint>`, and `<pp>`.

During production of the BNC, the n attribute was used with both `<author>` and `<imprint>` elements to supply a six-letter code identifying the author or imprint concerned. The values used should be unique across the corpus.

The `<imprint>` element is supplied for published texts only and contains the following elements in the order given: `<pubPlace>`, `<publisher>`, and `<date>`.

The following example demonstrates how these elements are used to record bibliographic details for a typical book:

```
<bibl>
 <title>It might have been Jerusalem.  </title>
 <author n="HealyT1" domicile="Scotland">Healy, Thomas</author>
 <imprint n="POLYGO1">
```

```
  <publisher>Polygon Books</publisher>
  <pubPlace>Edinburgh</pubPlace>
  <date value="1991">1991</date>
 </imprint>
 <pp>1-81</pp>
</bibl>
<!- BNC ->
```

The following example is typical of the case where a collection of leaflets or newsletters has been treated as a single text:

```
<bibl>
 <title>[Potato Marketing Board leaflets]</title>
 <imprint n="POTATO1">
  <publisher>Potato Marketing Board</publisher>
  <pubPlace>London</pubPlace>
  <date value="1991">1991</date>
 </imprint>
</bibl>
<!- EEA ->
```

6.1.6 The encoding description

The second major component of the TEI header is the encoding description (`<encodingDesc>`). This contains information about the relationship between an encoded text and its original source and describes the editorial and other principles employed throughout the corpus. It also contains reference information used throughout the corpus.

The BNC `<encodingDesc>` element has the following six components:

`<projectDesc>` describes in detail the purpose for which an electronic file was encoded, together with any other relevant information concerning the process by which it was assembled or collected.

`<samplingDecl>` contains a prose description of the rationale and methods used in sampling texts in the creation of the corpus.

`<editorialDecl>` provides details of editorial principles and practices applied during the encoding of a text.

`<tagsDecl>` provides detailed information about the tagging applied to a corpus text.

`<refsDecl>` specifies how canonical references are constructed for a text.

`<classDecl>` contains a series of `<category>` elements, defining the classification codes used for texts within the corpus.

One of each of these elements appears in the corpus header. Only the `<tagsDecl>` element appears in the individual text headers.

Documentary components of the encoding description

The `<projectDesc>` element for the corpus gives a brief description of the goals, organization and results of the BNC project. The `<samplingDecl>`, `<editorialDecl>` and `<refsDecl>` elements similarly supply brief prose descriptions describing the sampling procedures used in the project and the referencing system applied. This information is also summarized elsewhere in this documentation.

The tagging declaration

The tagging declaration (<tagsDecl>) element is used slightly differently in corpus and in text headers. In the corpus header, it is used to list every element name actually used within the corpus, together with a brief description of its function. In text headers, it is used to specify the number of elements actually tagged within each text. In either case it consists of a <namespace> element, containing a number of <tagUsage> elements, each of which has attributes gi and occurs.

In the corpus header, each <tagUsage> element contains a brief description of the element specified by its <gi> element; the occurs attribute is not supplied, as in the following extract:

```
<tagUsage gi="event">Non-verbal event in spoken text
</tagUsage>
<tagUsage gi="gap">Point where source material has omitted
</tagUsage>
<tagUsage gi="head">Header or headline in written text
</tagUsage>
```

In text headers, the <tagUsage> elements are empty, but the occurs attribute is always supplied, and indicates the number of such elements which appear within the text, as in the following example, taken from a typical written text:

```
<tagsDecl>
 <namespace name="">
  <tagUsage gi="c" occurs="5750"/>
  <tagUsage gi="corr" occurs="1"/>
  <tagUsage gi="div" occurs="115"/>
  <tagUsage gi="gap" occurs="3"/>
  <tagUsage gi="head" occurs="156"/>
  <tagUsage gi="hi" occurs="147"/>
  <tagUsage gi="l" occurs="2"/>
  <tagUsage gi="lg" occurs="1"/>
  <tagUsage gi="mw" occurs="256"/>
  <tagUsage gi="p" occurs="680"/>
  <tagUsage gi="quote" occurs="3"/>
  <tagUsage gi="s" occurs="2415"/>
  <tagUsage gi="w" occurs="41799"/>
 </namespace>
</tagsDecl>
```

The reference and classification declarations

The <refsDecl> element for the corpus header defines the approved format for references to the corpus. It takes the following form

```
<refsDecl>
 <para>Canonical references in the British National Corpus
   are to text segment (s) elements, and are constructed by taking the
value of the xml:id attribute of the bncDoc element containing the target
text, and concatenating a dot separator, followed by the value of the n
attribute of the target s element.</para>
</refsDecl>
```

The standard TEI <classDecl> element is used in the BNC Corpus Header to formally define several text classication schemes which are used in the corpus. Each scheme or taxonomy defines a number of code/description pairs, applicable to a text in the corpus. For example, the written domain taxonomy defines twelve subject domains ("Imagination", "Informative: natural science", "Informative:

applied science" etc.) and each written text is assigned to one of them. Each taxonomy is defined in the corpus header, using the following elements:

Here, for example, is the start of the `<taxonomy>` element defining the Written domain classification system as it appears in the corpus header:

```
<taxonomy xml:id="WRIDOM">
 <desc>Written Domain</desc>
 <category xml:id="WRIDOM1">
  <catDesc>Imaginative</catDesc>
 </category>
 <category xml:id="WRIDOM2">
  <catDesc>Informative:  natural & pure science</catDesc>
 </category>
 <category xml:id="WRIDOM3">
  <catDesc>Informative:  applied science</catDesc>
 </category>
...
</taxonomy>
```

A complete list of the taxonomies used in the BNC is included in the corpus header. The relevant chapter of the User Reference Guide specifies further how these categories are used in the full corpus; they are of less importance for the comparatively simple design of BNC Baby. The classification categories applicable to a given text are specified by the `<catRef>` element within the associated text header. Its target lists the identifiers of all `<category>` elements applicable to that text. For example, the header of a written text assigned to the social science domain which has a corporate author will include a `<catRef>` element like the following:

```
<catRef target="...  WRIATY1 WRIDOM4..."/>
```

(The dots above represent the identifiers of all other category codes applicable to this text).

Further information about the classification and categorization of an individual texts is provided within the `<textClass>` element discussed below (*6.1.7.5. Text classification*)

The Xaira Specification

The Xaira Specification element is used solely by the XAIRA indexing software. Its use in the BNC is further described in the full User Reference Guide. For more information, consult also the Xaira documentation available from http://www.xaira.org/

6.1.7 The profile description

The third component of a TEI header is the profile description. In the BNC this has the following components:

`<creation>` contains information about the creation of a text.

`<langUsage>` describes the languages, sublanguages, registers, dialects etc. represented within a text.

`<particDesc>` describes the identifiable participants in a linguistic interaction together with their relationships, where known.

`<settingDesc>` describes the setting or settings within which a language interaction takes place.

`<textClass>` groups information which describes the nature or topic of a text in terms of a standard classification scheme, thesaurus, etc.

The creation element

This element is provided to record the date of publication for texts originally published separately, and any details concerning the origination of any spoken or written texts, whether or not covered elsewhere. It is supplied in every text header, although the details provided vary. As a minimum, a date (tagged with the standard <date> element) will be included; this gives the date the content of this text was first created. For a spoken text, this will be the same as the date of the recording; for a written text, it will normally be the date of first publication of the edition, which may not be the same as the date of publication of the copy used.

Here are two typical examples:

```
<creation>
 <date>1992-02-11</date>:

</creation>
<creation>
 <date>1971</date>:  originally published by Jonathan Cape.

</creation>
```

Note that the BNC contains modernized editions of some classic texts such as Defoe's *Robinson Crusoe* (FRX); the creation date specified here is that of the creation of the modernized version rather than the 17th c. original.

For imaginative works, the creation date is also the date used to classify the text (by means of the WRITIM category). For other written works, such as textbooks, which are likely to have been extensively revised since their first publication, the date used to classify the text will be that of the edition described in the <sourceDesc>, but the original date will also be recorded within the <creation> element.

The <langUsage> element

Unlike the other elements of the profile description, the language usage element occurs only in the corpus header. It contains the following text:

```
<langUsage>
 <language ident="en-GB">The language of the British National Corpus is
modern British English.  Words,
   fragments, and passages from many other languages, both ancient and
modern, occur within the corpus where these may be represented using a
Latin alphabet.  Long passages in these languages, and material in other
languages, are generally silently deleted.</language>
</langUsage>
```

The participant description

The participant description (<particDesc>) element is used to provide information about speakers of texts transcribed for the BNC. It appears only within individual spoken text headers to define the participants specific to those texts.

It contains a series of <person> elements describing the participants whose speech is transcribed in this text.

The person element Each <person> element describes a single participant in a language interaction. It carries a number of attributes which are used to provide encoded values for some key aspects of the person concerned.

The xml:id attribute is required for each participant whose speech is included in a text, and its value is unique within the corpus. Although a given individual will always have the same identifier within a single text, there is no way of identifying the same individual should they appear in different texts. Since all demographically sampled conversations collected by a single respondent are treated together as a single text, and respondents were recruited from many different social contexts, the probability of the same person being recorded by different respondents is rather low, though not completely impossible.

On many occasions the speaker of a given utterance cannot be identified. A special code is used to indicate an unknown speaker, but, for consistency, this is also made unique to each text. Thus, an "unknown speaker" in one text will have different identifying code from an "unknown speaker" in another. As far as possible, different speakers are given different identifying codes, even where they cannot be identified with any confidence; thus there may be more than one "unidentified" speaker in the same text.

Where several speakers speak together, if they are identified, then all of the relevant codes are given; if however they are not, then a special "unknown speaker group" code is used.

The sex attribute specifies the sex of the participant, as one of female (284 cases in BNC Baby), male (214 cases in BNC Baby), or unknown (68).

The ageGroup attribute specifies the age group to which the participant belonged at the time of recording. The following breakdown applies to the participants in BNC Baby:

Table 10: Participant age groups

Code	Age Range	number such
Ag0	Under 15 years	47
Ag1	15 to 24 years	25
Ag2	25 to 34 years	41
Ag3	35 to 44 years	41
Ag4	45 to 59 years	35
Ag5	Over 59 years	29
X	Unknown	76

The class attribute specifies the social class to which the participant was assigned by the respondent. The following breakdown applies to the participants in BNC Baby:

Table 11: Participant social classes

Code	Social class	number such
UU	Unknown	388
AB	Managerial, administrative, professional	30
C1	Junior management, supervisory, professional	58
C2	Skilled manual	40
DE	Semi- or unskilled	14

The dialect attribute specifies a code for any region or dialect assigned to the participant's speech by the respondent. A full list of the codes used and their meanings is given in the User Reference Guide;. In BNC Baby, the most frequently occurring values, accounting for 480 of the 530 different codes used are

Table 12: Participant perceived dialect

Code	Region/dialect	number such
XOT	Unidentified dialect	142
NONE	No dialect	58
XME	North East Midlands	34
XWA	Welsh	32

Participant perceived dialect(cont.)

XLO	London	32
XSL	South West England	32
XSD	Scottish	30
XHC	Home Counties	28
XLC	Lancashire	18
XNO	Northern England	16
XNE	North East England	16
XMC	Central Midlands	14
XMS	South Midlands	8
XIR	Irish	8
XUR	European	8

Where it is available, additional information about a participant may be provided by one or more of the following elements, appearing within the `<person>` element: `<persName>` (personal name); `<occupation>`; `<dialect>`; `<age>` (age in years at time of recording); `<persNote>` (used for any other comment about the person recorded by the respondent).

In each case, the information provided is that given by the respondent and is taken from the log books issued to all participants in the demographic part of the corpus. It has not been normalized.

Here is a typical example from the demographic part of the corpus:

```
<person
   ageGroup="Ag0"
   dialect="XLO"
   xml:id="PS5A1"
   role="self"
   sex="m"
   soc="C2">
 <name>Terry</name>
 <age>14</age>
 <occupation>student</occupation>
 <dialect>London</dialect>
</person>
```

The role attribute is used to specify the role of this participant with respect to the respondent, as specified by the respondent. The value is not constrained; the most frequent cases found in BNC Baby are friend, self, other, colleague, son, husband, daughter, stranger, employee, wife, and mother. For example, the participant information recorded in the header for a text (KSU) comprising conversations between four participants: Michael and Steve (who are brothers), their mother Christine and their aunt Leslie is as follows:

```
<particDesc n="708">
 <person
    ageGroup="Ag0"
    xml:id="PS6RM"
    role="self"
    sex="m"
    soc="C2"
    dialect="XOT">
  <age>13</age>
  <persName>Michael</persName>
  <occupation>student</occupation>
 </person>
 <person
    ageGroup="Ag4"
```

```
   xml:id="PS6RN"
   role="mother"
   sex="f"
   soc="C2"
   dialect="XOT">
  <age>45</age>
  <persName>Christine</persName>
  <occupation>credit controller</occupation>
 </person>
 <person
   ageGroup="Ag4"
   xml:id="PS6RP"
   role="aunt"
   sex="f"
   soc="UU"
   dialect="XOT">
  <age>45</age>
  <persName>Leslie</persName>
  <occupation>unemployed</occupation>
 </person>
 <person
   ageGroup="Ag1"
   xml:id="PS6RR"
   role="brother"
   sex="m"
   soc="C2"
   dialect="XOT">
  <age>21</age>
  <persName>Steve</persName>
  <occupation>unemployed</occupation>
 </person>
</particDesc>
```

The setting description

The <settingDesc> element is used to describe the context within which a spoken text takes place. It appears once in the header of each spoken text, and contains one or more <setting> elements for each distinct recording. Its n attribute supplies the identifier of the tape recording made at this location; its who attribute supplies a list of identifiers for the person or persons recorded there.

The content of each <setting> element supplies additional details about the place (using <locale> and <placeName> elements); the time of day (using a <date> element); and about any other activities going on (using an <activity> element).

Some typical examples follow:

```
<setting n="020901" who="PS000 DCJPS000 DCJPS001">
 <name>Essex:  Harlow </name>
 <locale> Harlow College</locale>
 <activity> A level lecture </activity>
</setting>
<setting xml:id="KDFSE002" n="063505" who="PS0M6">
 <name>Lancashire:  Morecambe </name>
 <locale> at home </locale>
 <activity> watching television </activity>
</setting>
```

Text classification

The TEI provides a number of ways in which classification or text-type information may be specified for a text, grouped together within a <textClass> element, which appears once in the header of each text. Classifications may be represented using references to internally defined classications provided in the <classCode> element (such as the BNC classification scheme described in section *6.1.6.3. The reference and classification declarations*), by reference to some other predefined classification system, or by an open set of keywords. All three methods are used in the BNC, using the following elements:

A <catRef> element is provided in the header of each text. Its target attribute contains values for each of the classification codes defined in the corpus header. In each case, the classification code consists of a code used as the identifier of a <category> element within a <taxonomy> element defined in the corpus header. For example: ALLTIM1 indicates ‘dated 1960-1974’. A list of the values used is given in section «classcodes» below.

This taxonomy is that originally defined for selection and description of texts during the design of the corpus, as further discussed elsewhere. It is of course possible to classify the texts in many other ways, and no claim is made that this method is universally applicable or even generally useful, though it does serve to identify broadly distinct sub-parts of the corpus for investigation. The reader is also cautioned that, although an attempt has been made in the current edition of the corpus to correct the more egregious classification errors noted in the first edition, unquestionably many errors and inconsistencies remain. In particular, the categories WRILEV (perceived level of difficulty) and WRISTA (estimated circulation size) were incorrectly differentiated during the preparation of the corpus and cannot be relied on.

A <classCode> element is also provided for every text in the corpus. This contains the code assigned to the text in a genre-based analysis carried out at Lancaster University by David Lee since publication of the first edition of the BNC. Lee’s scheme classes the texts more delicately in most cases, since it takes into account their topic or subject matter (see further «classcodes» below).

Lee’s scheme is also used as the basis of a very simple categorization for each text, which is provided by means of the type attribute on its <text> or <stext> element. This categorization distinguishes six categories for written text (fiction, academic prose, non-academic prose, newspapers, other published, unpublished), and two for spoken text (conversation, other); It may be found a convenient way of distinguishing the major text types represented in the corpus: see further «tagcounts».

In the first release of the BNC, most texts were assigned a set of descriptive keywords, tagged as <term> elements within the <keywords> element. These terms were not taken from any particular descriptive thesaurus or closed vocabulary; the words or phrases used are those which seemed useful to the data preparation agency concerned, and are thus often inconsistent or even misleading. They have been retained unchanged in the present version of the BNC, pending a more thorough revision. In the World (second) Edition this set of keywords was complemented for most written texts by a second set, also tagged using a <keywords> element, but with a value for its source attribute of COPAC, indicating that the terms so tagged are derived from a different source. The source used was a major online library catalogue service (see http://www.copac.ac.uk). Like other public access catalogue systems, COPAC uses a well-defined controlled list of keywords for its subject indexing, details of which are not further given here.

Here is an example showing how one text (BND) is classified in each of these ways:

```
<teiHeader>...  <textClass>
  <catRef
    targets="WRI ALLTIM3 ALLAVA2 ALLTYP3 WRIAAG0 WRIAD0 WRIASE3 WRIATY2
WRIAUD3 WRIDOM8 WRILEV2 WRIMED1 WRIPP5 WRISAM1 WRISTA2 WRITAS3"/>
  <classCode scheme="DLee">W_religion</classCode>
  <keywords scheme="COPAC">
   <term>Marriage - Religious aspects - Christianity</term>
   <term>Marriage - Christian viewpoints</term>
  </keywords>
  <keywords>
   <term>Christian guide to marriage</term>
  </keywords>
 </textClass>...</teiHeader>
<wtext type="NONAC">...</wtext>
```

6.1.8 The revision description

The revision description (<revisionDesc>) element is the fourth and final element of a standard TEI header. In the BNC, it consists of a series of <change> elements.

Here is part of a typical example:

```
<revisionDesc>
 <change date="2006-10-21" who="#OUCS">Tag usage updated for
BNC-XML</change>
 <change date="2000-12-13" who="#OUCS">Last check for BNC World first
   release</change>
...<change date="1999-12-25" who="#OUCS">corrected tagUsage</change>
 <change date="1999-09-13" who="#UCREL">POS codes revised for BNC-2; header
   updated</change>
 <change date="1994-11-24" who="#dominic">Initial accession to
   corpus</change>
</revisionDesc>
```

7 Word class codes

As noted above, a full discussion of the principles and practice underlying the CLAWS word class annotation scheme used in all versions of the BNC is provided in the online User Reference Guide at http://www.natcorp.ox.ac.uk/XMLedition/URG/posguide.html. For convenience, a list of the codes used is also provided here.

POS	usage	POS	usage
AJ0	Adjective (general or positive) (e.g. *good*, *old*, *beautiful*)		
AJC	Comparative adjective (e.g. *better*, *older*)	AJS	Superlative adjective (e.g. *best*, *oldest*)
AT0	Article (e.g. *the*, *a*, *an*, *no*)	AV0	General adverb: an adverb not subclassified as AVP or AVQ (see below) (e.g. *often*, *well*, *longer* (adv.), *furthest*.
AVP	Adverb particle (e.g. *up*, *off*, *out*)	AVQ	Wh-adverb (e.g. *when*, *where*, *how*, *why*, *wherever*)
CJC	Coordinating conjunction (e.g. *and*, *or*, *but*)	CJS	Subordinating conjunction (e.g. *although*, *when*)
CJT	The subordinating conjunction *that*	CRD	Cardinal number (e.g. *one*, *3*, *fifty-five*, *3609*)
DPS	Possessive determiner-pronoun (e.g. *your*, *their*, *his*)	DT0	General determiner-pronoun: i.e. a determiner-pronoun which is not a DTQ or an AT0.
DTQ	Wh-determiner-pronoun (e.g. *which*, *what*, *whose*, *whichever*)	EX0	Existential there, i.e. *there* occurring in the *there is ...* or *there are ...* construction
ITJ	Interjection or other isolate (e.g. *oh*, *yes*, *mhm*, *wow*)	NN0	Common noun, neutral for number (e.g. *aircraft*, *data*, *committee*)
NN1	Singular common noun (e.g. *pencil*, *goose*, *time*, *revelation*)	NN2	Plural common noun (e.g. *pencils*, *geese*, *times*, *revelations*)
NP0	Proper noun (e.g. *London*, *Michael*, *Mars*, *IBM*)	ORD	Ordinal numeral (e.g. *first*, *sixth*, *77th*, *last*) .
PNI	Indefinite pronoun (e.g. *none*, *everything*, *one* [as pronoun], *nobody*)	PNP	Personal pronoun (e.g. *I*, *you*, *them*, *ours*)

`PNQ`	Wh-pronoun (e.g. *who*, *whoever*, *whom*)	`PNX`	Reflexive pronoun (e.g. *myself*, *yourself*, *itself*, *ourselves*)
`POS`	The possessive or genitive marker *’s* or *’*	`PRF`	The preposition *of*
`PRP`	Preposition (except for *of*) (e.g. *about*, *at*, *in*, *on*, *on behalf of*, *with*)	`PUL`	Punctuation: left bracket - i.e. *(* or *[*
`PUN`	Punctuation: general separating mark - i.e. *.* *,* *!* *,* *:* *;* - or *?*	`PUQ`	Punctuation: quotation mark - i.e. *’* or *"*
`PUR`	Punctuation: right bracket - i.e. *)* or *]*	`TO0`	Infinitive marker *to*
`UNC`	Unclassified items which are not appropriately considered as items of the English lexicon.	`VBB`	The present tense forms of the verb BE, except for *is*, *’s*: i.e. *am*, *are*, *’m*, *’re* and *be* [subjunctive or imperative]
`VBD`	The past tense forms of the verb BE: *was* and *were*	`VBG`	The -ing form of the verb BE: *being*
`VBI`	The infinitive form of the verb BE: *be*	`VBN`	The past participle form of the verb BE: *been*
`VBZ`	The -s form of the verb BE: *is*, *’s*	`VDB`	The finite base form of the verb BE: *do*
`VDD`	The past tense form of the verb DO: *did*	`VDG`	The -ing form of the verb DO: *doing*
`VDI`	The infinitive form of the verb DO: *do*	`VDN`	The past participle form of the verb DO: *done*
`VDZ`	The -s form of the verb DO: *does*, *’s*	`VHB`	The finite base form of the verb HAVE: *have*, *’ve*
`VHD`	The past tense form of the verb HAVE: *had*, *’d*	`VHG`	The -ing form of the verb HAVE: *having*
`VHI`	The infinitive form of the verb HAVE: *have*	`VHN`	The past participle form of the verb HAVE: *had*
`VHZ`	The -s form of the verb HAVE: *has*, *’s*	`VM0`	Modal auxiliary verb (e.g. *will*, *would*, *can*, *could*, *’ll*, *’d*)
`VVB`	The finite base form of lexical verbs (e.g. *forget*, *send*, *live*, *return*) [Including the imperative and present subjunctive]	`VVD`	The past tense form of lexical verbs (e.g. *forgot*, *sent*, *lived*, *returned*)
`VVG`	The -ing form of lexical verbs (e.g. *forgetting*, *sending*, *living*, *returning*)	`VVI`	The infinitive form of lexical verbs (e.g. *forget*, *send*, *live*, *return*)
`VVN`	The past participle form of lexical verbs (e.g. *forgotten*, *sent*, *lived*, *returned*)	`VVZ`	The -s form of lexical verbs (e.g. *forgets*, *sends*, *lives*, *returns*)
`XX0`	The negative particle *not* or *n’t*	`ZZ0`	Alphabetical symbols (e.g. *A*, *a*, *B*, *b*, *c*, *d*)

In addition to the basic 57 codes tabulated above, the BNC World Edition uses thirty ‘portmanteau’ or ‘ambiguity’ tags. These are applied wherever the probabilities assigned by the CLAWS automatic tagger to its first and second choice tags were considered too low for reliable disambiguation. So, for example, the ambiguity tag `AJ0-AV0` indicates that the choice between adjective (`AJ0`) and adverb (`AV0`) is left open, although the tagger has a preference for an adjective reading. The mirror tag, `AV0-AJ0`, again shows adjective-adverb ambiguity, but this time the more likely reading is the adverb.

This table lists, for each of the twelve simplified wordclass tags used by the `pos` attribute, the corresponding CLAWS C5 tags of which the class consists.

POS	significance	combines

ADJ	adjective	AJ0, AJC, AJS, CRD, DT0, ORD
ADV	adverb	AV0, AVP, AVQ, XX0
ART	article	AT0
CONJ	conjunction	CJC, CJS, CJT
INTERJ	interjection	ITJ
PREP	preposition	PRF, PRP, TO0
PRON	pronoun	DPS, DTQ, EX0, PNI, PNP, PNQ, PNX
STOP	punctuation	POS, PUL, PUN, PUQ, PUR
SUBST	substantive	NN0, NN1, NN2, NP0, ONE, ZZ0, NN1-NP0, NP0-NN1
UNC	unclassified, uncertain, or non-lexical word	UNC, AJ0-AV0, AV0-AJ0, AJ0-NN1, NN1-AJ0, AJ0-VVD, VVD-AJ0, AJ0-VVG, VVG-AJ0, AJ0-VVN, VVN-AJ0, AVP-PRP, PRP-AVP, AVQ-CJS, CJS-AVQ, CJS-PRP, PRP-CJS, CJT-DT0, DT0-CJT, CRD-PNI, PNI-CRD, NN1-VVB, VVB-NN1, NN1-VVG, VVG-NN1, NN2-VVZ, VVZ-NN2
VERB	verb	VBB, VBD, VBG, VBI, VBN, VBZ, VDB, VDD, VDG, VDI, VDN, VDZ, VHB, VHD, VHG, VHI, VHN, VHZ, VM0, VVB, VVD, VVG, VVI, VVN, VVZ, VVD-VVN, VVN-VVD

8 Software for BNC Baby

8.1 Why XML?

A design goal of the original BNC project was that it should not be delivered in a format which was proprietary or which required the use of any particular piece of software. This, together with the desire to conform to emerging international standards, was a key factor in determining the choice of SGML as the vehicle for the corpus interchange format. A decade after this decision, SGML is still a widely used international standard format for which many public domain and commercial utilities exist. In the shape of XML, which is a simplified version of the original standard, SGML now dominates development of the world wide web, and hence of most sectors of the information processing community. New XML software appears almost every week, and it has been adopted by all current ‘major players’ from Sun and IBM to Microsoft.

That said, it must be recognised that the requirements of corpus linguists and others wishing to make use of the BNC are often rather specialised, and therefore unlikely to be supported by mainstream commercially produced software. For this and other reasons, the research user of the BNC who wishes to go beyond simple concordancing or word searching activities should expect to have to do some programming. This is another reason behind the choice of XML as a vehicle for the system: because of the wide take up of this language, there exist many utility libraries and generic programming interfaces which greatly simplify such processes as extracting the tags from a file, selecting portions of the text according to its logical structure, picking out files with certain attributes by searching their headers, and so on.

The BNC uses XML in a simple and straightforward way described in the rest of this manual; simple programs can be readily written using standard UNIX utilities such as grep or perl to access the corpus just as plain text files. More reliably, XML-aware programs can be written to application programming interfaces (APIs) such as the W3C’s Document Object Model (DOM) or the Simple API for XML (SAX), using application libraries developed for almost every modern programming or scripting language (C, Perl, Python, tcl, PHP etc.). Furthermore, a standard stylesheet language (XSLT) now exists which can be used to specify the transformation of XML source texts into almost any desired format by those with only minimal programming skills. Information about software resources is not provided here, but is readily found on the World Wide Web: currently, one good place to start looking is www.xml.com. Increasingly, support for XML is built into standard utilities such as web browsers, database systems, and stylesheet processors offering a high level of sophistication are readily available.

The XML edition of the BNC, and therefore also BNC Baby, is also provided along with its own access software called XAIRA (which can, incidentally, be used with any collection of XML texts, not simply the BNC). It should be emphasized however that use of the BNC is not synonymous with use of XAIRA. Most generic tools developed for corpus linguistics and NLP can be used with the BNC, although the tools may be vary in the extent to which they can make use of the markup in the corpus.[3]

Whatever software is used, it is necessary to have a clear understanding of the various elements tagged in the corpus, the contexts in which they may appear, and their intended semantics. The syntax of an XML document is defined by a schema. For TEI conformant texts, the TEI Header provides additional meta-information. The semantics of XML elements are provided by documentation such as that provided elsewhere in this manual.

8.2 BNC Baby delivery format

The corpus is delivered as a collection of 128 individual text files, grouped into four subdirectories, one for each of the text registers making up the corpus (aca for academic prose; fic for fictional texts; dem for demographically-sampled spoken texts; news for newspapers). The files are named in exactly the same way as in the BNC proper. Each file contains a single BNC document, i.e. a TEI header and its associated spoken or written text, and has the same name as the value of the id attribute on its `<bncDoc>` element, for example ABC.xml.

Note that the three-character identifiers used (and hence the filenames) are *entirely arbitrary* and do not convey any information about the type of text contained. Each text contains a TEI Header which specifies all such meta information, either directly, or by reference to the corpus header, as described in section *6. The header*. The corpus header is provided in a file called bncHdr.xml; it may be needed for XML processing is to be applied across the whole corpus, because it contains declarations of elements (such as the classification records) referred to by almost every part of the corpus.

8.3 XML processing

Some ancillary files relating to the encoding and processing of the corpus are included in the standard release in a subdirectory called XML. This contains two subdirectories:

scripts contains example XSLT scripts for processing BNC files

schemas contains versions of the BNC XML schema, expressed in RELAX NG compact or full syntax, W3C Schema Language, XML DTD, and TEI ODD languages.

The scripts directory includes the following simple demonstration XSLT stylesheets:

display.xsl converts a BNC text to an HTML format which can be read directly in a browser

justTheText.xsl removes all the tagging from a BNC text; also removes the whole of the header.

oneWordPerLine.xsl converts a BNC text to a "one word per line" format

justTheCodes.xsl removes all the words from a BNC text; also removes the whole of the header.

whoSpeaks.xsl produces a "dramatis personae" for a spoken text indicating how much each speaker contributes to it.

These stylesheets are for demonstration purposes only. They can be used by a standalone XSLT processor such as xsltproc, or referenced by any XSLT-aware web browser (such as Internet Explorer or Opera) or XML editor (such as Oxygen).

For example, to produce an HTML version of text ABC using the xsltproc utility, one might type

[3] Recent examples of concordancing tools which have been used successfully with the BNC include Xaira, BNCWeb, WordSmith, and the Sketch Engine

```
xsltproc XML/Scripts/display.xsl Texts/A/AB/ABC.xml > ABC.html
```

at the command prompt. Alternatively, to render the file directly on screen using Internet Explorer, one might add a stylesheet instruction like the following at the start of the file:

```
<?xml-stylesheet type="text/xsl"
href="/fullPathTo/XML/scripts/display.xsl" ?>
```

and then simply open the resulting file with the browser.

The BNC XML schema, in whatever form, is primarily useful as a means of validating and documenting the structure of the corpus files. For some purposes, chiefly the validation of the classification codes used, it may be necessary to process the corpus header along with the individual text or texts to be processed. For most purposes, however, individual texts in the corpus can be regarded as free-standing.

9 Lists of works excerpted

Brief bibliographic details of all the texts included in BNC Baby are given here, in alphabetical order of their identifier.

[A1E] 9537 words from *Independent, electronic edition of 1989-10-02: Business section.* Newspaper Publishing plc London 1989

[A1F] 8717 words from *Independent, electronic edition of 1989-10-02: Editorials* Newspaper Publishing plc London 1989

[A1G] 9909 words from *Independent, electronic edition of 1989-10-02: Foreign news pages.* Newspaper Publishing plc London 1989

[A1H] 3007 words from *Independent, electronic edition of 1989-10-02: Gazette section.* Newspaper Publishing plc London 1989

[A1J] 13407 words from *Independent, electronic edition of 1989-10-02: Home news pages.* Newspaper Publishing plc London 1989

[A1K] 1842 words from *Independent, electronic edition of 1989-10-02: Listings section.* Newspaper Publishing plc London 1989

[A1L] 1788 words from *Independent, electronic edition of 1989-10-02: Living section.* Newspaper Publishing plc London 1989

[A1M] 4808 words from *Independent, electronic edition of 1989-10-02: Science pages.* Newspaper Publishing plc London 1989

[A1N] 14330 words from *Independent, electronic edition of 1989-10-02: Sport section.* Newspaper Publishing plc London 1989

[A1P] 2488 words from *Independent, electronic edition of 1989-10-02: Frontpages.* Newspaper Publishing plc London 1989

[A1U] 4059 words from *Independent, electronic edition of 1989-10-03: Focus section.* Newspaper Publishing plc London 1989

[A1X] 3223 words from *Independent, electronic edition of 1989-10-03: Health pages.* Newspaper Publishing plc London 1989

[A2D] 2632 words from *Independent, electronic edition of 1989-10-04: Media section.* Newspaper Publishing plc London 1989

[A31] 3420 words from *Independent, electronic edition of 1989-10-06: Law section.* Newspaper Publishing plc London 1989

[A36] 5840 words from *Independent, electronic edition of 1989-10-07: Book reviews.* Newspaper Publishing plc London 1989

[A38] 3127 words from *Independent, electronic edition of 1989-10-07: Motoring pages.* Newspaper Publishing plc London 1989

[A39] 2234 words from *Independent, electronic edition of 1989-10-07: Antiques pages.* Newspaper Publishing plc London 1989

[A3C] 8234 words from *Independent, electronic edition of 1989-10-07: Food pages.* Newspaper Publishing plc London 1989

[A3E] 1831 words from *Independent, electronic edition of 1989-10-07: Gardening pages.* Newspaper Publishing plc London 1989

[A3K] 3396 words from *Independent, electronic edition of 1989-10-07: Property section.* Newspaper Publishing plc London 1989

[A3M] 2875 words from *Independent, electronic edition of 1989-10-07: Style section.* Newspaper Publishing plc London 1989

[A3P] 7721 words from *Independent, electronic edition of 1989-10-07: Travel section.* Newspaper Publishing plc London 1989

[A4D] 3051 words from *Independent, electronic edition of 1989-10-11: Architecture section.* Newspaper Publishing plc London 1989

[A5E] 5206 words from *Independent, electronic edition of 1989-10-14: Arts section.* Newspaper Publishing plc London 1989

[A6U] 26310 words from *Oxford Art Journal.* Oxford University Press Oxford 1991

[A7S] 5263 words from *The Guardian, electronic edition of 1989-11-08: Arts section.* Guardian Newspapers Ltd London 1989

[A7T] 8406 words from *The Guardian, electronic edition of 1989-11-08: Business section.* Guardian Newspapers Ltd London 1989

[A7W] 24367 words from *The Guardian, electronic edition of 1989-11-08: Home news pages.* Guardian Newspapers Ltd London 1989

[A7X] 711 words from *The Guardian, electronic edition of 1989-11-08: Leisure pages.* Guardian Newspapers Ltd London 1989

[A7Y] 10463 words from *The Guardian, electronic edition of 1989-11-08: News and features.* Guardian Newspapers Ltd London 1989

[A80] 10300 words from *The Guardian, electronic edition of 1989-11-08: Sport section.* Guardian Newspapers Ltd London 1989

[A82] 218 words from *The Guardian, electronic edition of 1989-11-11: Applied Science pages.* Guardian Newspapers Ltd London 1989

[A84] 390 words from *The Guardian, electronic edition of 1989-11-11: Religious affairs stories.* Guardian Newspapers Ltd London 1989

[A8L] 1032 words from *The Guardian, electronic edition of 1989-11-23: Leisure pages.* Guardian Newspapers Ltd London 1989

[A8M] 3453 words from *The Guardian, electronic edition of 1989-11-23: News and features.* Guardian Newspapers Ltd London 1989

[A8N] 11622 words from *The Guardian, electronic edition of 1989-11-23: Sport section.* Guardian Newspapers Ltd London 1989

[A8P] 2346 words from *The Guardian, electronic edition of 1989-11-23: Foreign news pages.* Guardian Newspapers Ltd London 1989

[A8R] 6474 words from *The Guardian, electronic edition of 1989-12-07: Applied Science pages.* Guardian Newspapers Ltd London 1989

[A8S] 1577 words from *The Guardian, electronic edition of 1989-12-07: Arts section.* Guardian Newspapers Ltd London 1989

[A8T] 660 words from *The Guardian, electronic edition of 1989-12-07: Religious affairs stories.* Guardian Newspapers Ltd London 1989

[A8U] 8415 words from *The Guardian, electronic edition of 1989-12-07: Business section.* Guardian Newspapers Ltd London 1989

[A91] 2333 words from *The Guardian, electronic edition of 1989-12-07: Foreign news pages.* Guardian Newspapers Ltd London 1989

[A97] 832 words from *The Guardian, electronic edition of 1989-12-08: Leisure pages.* Guardian Newspapers Ltd London 1989

[A98] 6610 words from *The Guardian, electronic edition of 1989-12-08: News and features.* Guardian Newspapers Ltd London 1989

[A9G] 280 words from *The Guardian, electronic edition of 1989-12-10: Leisure pages.* Guardian Newspapers Ltd London 1989

[A9J] 3526 words from *The Guardian, electronic edition of 1989-12-10: Foreign news pages.* Guardian Newspapers Ltd London 1989

[A9P] 576 words from *The Guardian, electronic edition of 1989-12-11: Leisure pages.* Guardian Newspapers Ltd London 1989

[A9X] 689 words from *The Guardian, electronic edition of 1989-12-13: Leisure pages.* Guardian Newspapers Ltd London 1989

[A9Y] 251 words from *The Guardian, electronic edition of 1989-12-13: News and features.* Guardian Newspapers Ltd London 1989

[AA3] 8674 words from *The Guardian, electronic edition of 1989-12-20: Business section.* Guardian Newspapers Ltd London 1989

[AA6] 1266 words from *The Guardian, electronic edition of 1989-12-20: Leisure pages.* Guardian Newspapers Ltd London 1989

[AAM] 1840 words from *The Guardian, electronic edition of 1989-12-22: Leisure pages.* Guardian Newspapers Ltd London 1989

[AAR] 2390 words from *The Guardian, electronic edition of 1989-12-31: Arts section.* Guardian Newspapers Ltd London 1989

[AB9] 40048 words from *Death of a partner.* Neel, Janet Constable & Company Ltd London 1991

[AC2] 35678 words from *Man at the sharp end.* Kilby, M The Book Guild Ltd Lewes, East Sussex 1991

[ACJ] 36837 words from *Principles of criminal law.* Ashworth, Andrew Oxford University Press Oxford 1991

[AHB] 16330 words from *Daily Telegraph, electronic edition of 1992-04-04: Business section.* The Daily Telegraph plc London 1992

[AHC] 38061 words from *Daily Telegraph, electronic edition of 1992-04-04: Leisure pages.* The Daily Telegraph plc London 1992

[AHD] 4120 words from *Daily Telegraph, electronic edition of 1992-04-04: Applied Science pages.* The Daily Telegraph plc London 1992

[AHE] 1198 words from *Daily Telegraph, electronic edition of 1992-04-04: News and features.* The Daily Telegraph plc London 1992

[AHF] 26386 words from *Daily Telegraph, electronic edition of 1992-04-04: Foreign news pages.* The Daily Telegraph plc London 1992

[AHH] 97 words from *Daily Telegraph, electronic edition of 1992-04-05: Religious affairs stories.* The Daily Telegraph plc London 1992

[AHL] 2476 words from *Daily Telegraph, electronic edition of 1992-04-05: Applied Science pages.* The Daily Telegraph plc London 1992

[AJ1] 387 words from *Daily Telegraph, electronic edition of 1992-04-07: Religious affairs stories.* The Daily Telegraph plc London 1992

[AJF] 6294 words from *Daily Telegraph, electronic edition of 1992-04-09: Arts section.* The Daily Telegraph plc London 1992

[AJG] 569 words from *Daily Telegraph, electronic edition of 1992-04-09: Religious affairs stories.* The Daily Telegraph plc London 1992

[AJW] 836 words from *Daily Telegraph, electronic edition of 1992-04-11: Religious affairs stories.* The Daily Telegraph plc London 1992

[AL0] 4958 words from *Daily Telegraph, electronic edition of 1992-04-17: Arts section.* The Daily Telegraph plc London 1992

[AL2] 8841 words from *Daily Telegraph, electronic edition of 1992-04-17: Business section.* The Daily Telegraph plc London 1992

[AL5] 2439 words from *Daily Telegraph, electronic edition of 1992-04-17: News and features.* The Daily Telegraph plc London 1992

[ALP] 25204 words from *British journal of social work.* Oxford University Press Oxford 1991

[AMM] 39025 words from *Fossils: the key to the past.* Fortey, Richard Natural History Museum Publications London 1991 52-123

[AS6] 29655 words from *Tackling the inner cities.* Pimlott, Ben MacGregor, Susanne Oxford University Press Oxford 1991 1-92

[B17] 33296 words from *Crime.* Marsh, Ian Longman Group UK Ltd Harlow 1992 1-94

[B1G] 37552 words from *Handling geographical information.* Blakemore, Michael Masser, Ian Longman Scientific & Technical Harlow 1991 55-176

[B2K] 24252 words from *Colliding plane waves in general relativity.* Griffiths, J B Oxford University Press Oxford 1991 48-145

[BM4] 54452 words from *The Alton Herald.* Farnham Castle Newspapers Ltd Farnham, Surrey 1992-03-27; 1992-04-1

[BMW] 40471 words from *Folly's child.* Tanner, Janet Century Hutchinson London 1991 13-124

[BPA] 36049 words from *The titron madness.* Bedford, John Dales Large Print Long Preston, N. Yorks 1984 11-209

[C8T] 39712 words from *Devices and desires.* James, P D Faber & Faber Ltd London 1989 17-120

[CB5] 39092 words from *Ruth Appleby.* Rhodes, Elvi Corgi Books London 1992 109-226

[CBD] 21252 words from *Today.* News Group Newspapers Ltd London 1992-12

[CBE] 44644 words from *Today.* News Group Newspapers Ltd London 1992-12

[CBM] 16128 words from *Ulster Newsletter.* Century Newspapers Ltd Belfast 1991-11-05

[CCW] 37971 words from *Crackdown.* Cornwell, Bernard Michael Joseph Ltd London 1990 15-124

[CDB] 37148 words from *A fatal inversion.* Vine, Barbara Viking London 1987 19-117

[CEL] 20709 words from *Today.* News Group Newspapers Ltd London 1992-12

[CFC] 42067 words from *East Anglian Daily Times.* East Anglian Daily Times Company Ipswich 1993-03

[CFY] 36220 words from *My beloved son.* Cookson, Catherine Corgi Books London 1992 85-221

[CH3] 117997 words from *The Daily Mirror.* Mirror Group Newspapers London 1992-08/1992-10

[CLP] 39995 words from *The mind at work.* Singleton, W T Cambridge University Press Cambridge 1989 1-129

[CLW] 37432 words from *Frameworks for teaching.* Dale, Roger Robinson, Alison Fergusson, R Hodder & Stoughton Ltd Sevenoaks, Kent 1992 201-299

[CMA] 39392 words from *Evolution from molecules to men.* Bendall, D S Cambridge University Press Cambridge 1985 43-565

[CRS] 39030 words from *Policies for diversity in education.* Swann, W Booth, T Masterton, M Potts, P Routledge & Kegan Paul plc London 1992 112-209

[CTY] 41600 words from *White mythologies: writing history and the West.* Young, R Routledge & Kegan Paul plc London 1990 1-90

[E9S] 15309 words from *The East Anglian.* East Anglian Daily Times Company Ipswich 1993-03

[EA7] 25172 words from *France in the making, 843-1180.* Dunbabin, Jean Oxford University Press Oxford 1991 223-335

[ECV] 39339 words from *Feminist perspectives in philosophy.* Whitford, M Griffiths, M Macmillan Publishers Ltd Basingstoke 1989 1-109

[EW1] 40989 words from *The age of Balfour and Baldwin 1902-1940.* Ramsden, John Longman Group UK Ltd Harlow 1978 65-151

[EWW] 37928 words from *Matrices and engineering dynamics.* Simpson, A Collar, A R Ellis Horwood Ltd Chichester 1987 11-195

[F98] 5602 words from *Computers and the humanities.* Kenny, A u.p.

[F9V] 38105 words from *Discourse analysis.* Yule, G Brown, G Cambridge University Press Cambridge 1991 1-112

[FAJ] 41125 words from *Masai dreaming.* Cartwright, J Macmillan Publishers Ltd Basingstoke 1993 50-182

[FC1] 2296 words from *The Weekly Law Reports 1992 Volume 3.* u.p.

[FEF] 25873 words from *Lectures on electromagnetic theory.* Solymar, L Oxford University Press Oxford 1984 5-118

[FET] 34083 words from *Still life.* Byatt, A S Penguin Group London 1988 1-84

[FPB] 39092 words from *Crimson.* Conran, S Penguin Group London 1992 289-426

[FPG] 38231 words from *Design of computer data files.* Hanson, O Pitman Publishing London 1989 144-287

[FSS] 42892 words from *The law of Tort.* Rogers, W V H Sweet & Maxwell Ltd London 1991 421-535

[FT1] 45374 words from *British Medical Journal.* British Medical Association London 1977 3-574

[G01] 40226 words from *Captives.* Hutson, Shaun Warner Books London 1992 163-322

[G0L] 40217 words from *The Lucy ghosts.* Shah, Eddy Corgi Books London 1993 321-452

[G0S] 41733 words from *Indigo.* Warner, Marina Chatto & Windus Ltd London 1992 143-268

[G0Y] 41526 words from *Passing on.* Lively, P Penguin Group London 1990 58-178

[GUU] 40020 words from *Freelance death.* Taylor, Andrew Victor Gollancz Ltd London 1993 52-175

[GVL] 41954 words from *The night mayor.* Newman, Kim New English Library Sevenoaks 1990 49-185

[H85] 41106 words from *Coffin underground.* Butler, Gwendoline Fontana Press London 1991 54-215

[H9C] 44223 words from *The prince of darkness.* Doherty, P C Headline Book Publishing plc London 1992

[H9D] 42267 words from *Posthumous papers.* Barnard, Robert Corgi Books London 1992 36-171

[HR9] 41905 words from *They came from SW19.* Williams, Nigel Faber & Faber Ltd London 1992 119-250

[HRG] 42483 words from *Polymers: the chemistry and physics of modern materials.* Cowie, J M G Blackie academic & professional Glasgow 1991 157-303

[HWV] 44333 words from *The Lancet.* The Lancet Ltd London 1993-06/19

[HXH] 44848 words from *How to write essays, dissertations & theses in literary studies.* Fabb, Nigel Durant, Alan Longman Group UK Ltd Harlow 1993 9-140

[J10] 42229 words from *The Mamur Zapt and the girl in the Nile.* Pearce, Michael Fontana Press London 1992 5-150

[J18] 44922 words from *Tropical rainforest ecology.* Mabberley, D J Blackie & Son Ltd London 1992 133-241

[J54] 35014 words from *The divided house.* Raymond, Mary F A Thorpe (Publishing) Ltd UK 1985 1-236

[J57] 36452 words from *The British polity.* Norton, Philip Longman New York 1984 3-101

[J7G] 4955 words from *Language and Literature Volume 1 No.2 1992.* Green, Keith Longman Group UK Ltd Harlow 1992

[K29] 1976 words from *Belfast Telegraph: Applied Science pages.* u.p.

[K2A] 7718 words from *Belfast Telegraph: Arts section.* u.p.

[K2B] 2104 words from *Belfast Telegraph: Religious affairs stories.* u.p.

[K2C] 3333 words from *Belfast Telegraph: Business section.* u.p.

[K2E] 5439 words from *Belfast Telegraph: News and features.* u.p.

[K2N] 20695 words from *Belfast Telegraph: Foreign news pages.* u.p.

[K36] 653 words from *Liverpool Daily Post and Echo]* u.p.

[K37] 8232 words from *Liverpool Daily Post and Echo: Arts section.* u.p.

[K38] 1250 words from *Liverpool Daily Post and Echo: Religious affairs stories.* u.p.

[K39] 6969 words from *Liverpool Daily Post and Echo: Business section.* u.p.

[K3A] 25658 words from *Liverpool Daily Post and Echo: Leisure pages.* u.p.

[K3B] 1369 words from *Liverpool Daily Post and Echo: News and features.* u.p.

[K3C] 36979 words from *Liverpool Daily Post and Echo: Foreign news pages.* u.p.

[K3D] 565 words from *Liverpool Daily Post and Echo: Applied Science pages.* u.p.

[K4R] 1084 words from *Northern Echo: Religious affairs stories.* u.p.

[K4S] 35935 words from *Northern Echo: Business section.* u.p.

[K4U] 1972 words from *Northern Echo: Applied Science pages.* u.p.

[K4Y] 26999 words from *Northern Echo: Arts section.* u.p.

[K58] 3874 words from *The Scotsman: Religious affairs stories.* u.p.

[K5B] 362 words from *The Scotsman: Applied Science pages.* u.p.

[K5C] 52665 words from *The Scotsman: News and features.* u.p.

[K5E] 7044 words from *The Scotsman: Applied Science pages.* u.p.

[K8V] 38611 words from *Coffin in fashion.* Butler, Gwendoline Fontana Press London 1990 61-221

[KB5] 5541 words from *11 conversations recorded by 'Andrew' (PS028, R 502) on 30 May 1991 with 6 interlocutors*

PS028 16, Andrew, student, London, self

PS0Y5 17, Suzanne, student, London, girlfriend

PS0Y7 46, Val, systems analyst, Northern England, mother

PS0Y8 17, John, student, London, friend

PS0YA 18, Sanjay, student, London, friend

PS0YC 18, Jon, student, London, friend

KB5PSUNK Unknown speaker, other

KB5PSUGP Group of unknown speakers, other

[KB7] 103997 words from *60 conversations recorded by 'Ann' (PS02G, R 28) between 28 November and 5 December 1991 with 35 interlocutors*

PS02G 46, Ann, team leader, North-east England, self

PS02H 33, Stuart, factory operative, East Anglia, partner

PS02J 45, Brian, paper-man, Northern England, stranger

PS02K 40, Jill, estate agent, East Anglia, stranger

PS02L 63, Jeanie, retired, Scottish, partner's mother

PS02M 40, Sue, care assistant, Scottish, stranger

PS02N 45, Shirley, publican, Midlands, friend

PS02P 50, Alan, Scottish, stranger

PS02R Terry, Scottish, stranger

PS02S 46, Bob, publican, Central Midlands, friend

PS02T 50, Harry, friend

PS02U 60, John, retired (policeman), London, friend

PS02V 23, Martin, marketing person, Northern England, friend

PS02X 22, Helen, housewife, East Anglia, daughter

PS02Y 17, Jonathan, Youth Training Scheme trainee, East Anglia, son

PS030 19, Christian, process worker, colleague

PS031 2, Richard, pre-school, grandson

PS032 1, Anna Marie, pre-school, granddaughter

PS033 26, Bernie, building labourer, son-in-law

PS036 61, Bill, team leader, colleague

PS037 57, Jean, process worker, colleague

PS038 26, Steven, process worker, colleague

PS039 44, Deidre, senior team leader, colleague

PS03A 56, John, process worker, colleague

PS03B 50, Dawn, process worker, colleague

PS03C 27, Tina, team leader, East Anglia, colleague

PS03D 36, Val, Queen's counsel, colleague

PS03F 40, Trevor, process worker, colleague

PS03H 44, Brenda, friend

PS03J 21, Lynda, friend

PS03K 50, Neville, friend

PS03M 25, Maxine, process worker, Northern England, colleague

PS03N 36, Mervin, process worker, East Anglia, colleague

PS03R Grace, ?

PS6TB None, ?

[KBC] 31337 words from *14 conversations recorded by 'Audrey' (PS1A9, R 606) between 2 and 9 April 1992 with 9 interlocutors*

PS1A9 61, Audrey, housewife, Lancashire, self

PS1AA 61, Gordon, teacher, Lancashire, husband

PS1AB 45, Margaret, nurse, Lancashire, friend

PS1AC 50+, Joan, clerk, Central Northern England, acquaintance

PS1AD 29, Kevin, computer engineer, Northern England, son-in-law

PS1AE 31, Carl, pharmacist, Northern England, son

PS1AF None, ?

PS1AG 28, Elaine, housewife, Northern England, daughter

PS1AH 60, Iris, housewife, Lancashire, cousin

KBCPSUNK Unknown speaker, other

KBCPSUGP Group of unknown speakers, other

[KBD] 58087 words from *24 conversations recorded by 'Barry' (PS03W, R 66) between 1 and 6 February 1992 with 10 interlocutors*

PS03W 41, Barry, entertainments consultant, Central Northern England, self

PS03X 35, Terri, bar staff, Home Counties, employee

PS03Y 30, Hugh, bar staff, Irish, employee

PS040 38, Alan, security, Lancashire, employee

PS041 None, ?

PS042 30, Mark, dj, London, employee

PS043 30, Ken, security, Lancashire, employee

PS044 35, None, housewife, Lancashire, employee's wife

PS045 9, Sergio, student (state primary), Lancashire, son

KBDPSUNK Unknown speaker, other

KBDPSUGP Group of unknown speakers, other

[KBH] 47995 words from *63 conversations recorded by 'Carol' (PS05B, R 104) between 30 May and 5 June 1991 with 8 interlocutors*

PS05B 36, Carole, administrator (pt), Home Counties, self

PS05C 18, Joelle, au pair, European (French), employee

PS05D 36, Adam, textile agent, Home Counties, husband

PS05E 2, Charlotte, pre-school, daughter

PS05F 30+, Pauline, shop owner, Home Counties, friend

PS05H 60+, Ron, retired, Home Counties, father

PS05J 60+, Pat, retired, Home Counties, mother

PS05K 5, Toby, nephew

KBHPSUNK Unknown speaker, other

KBHPSUGP Group of unknown speakers, other

[KBJ] 11137 words from *26 conversations recorded by 'Carol2' (PS1DP, R 12) between 10 and 17 January 1992 with 6 interlocutors*

PS1DP 44, Carol, courier, Central Midlands, self

PS1DR 11, Emma, student (state secondary), Central Midlands, daughter

PS1DS 47, Chris, area organiser for union, Central Midlands, husband

PS1DT 34, Dave, miner, Central Midlands, friend

PS1DU 26, Ann, hairdresser, Central Midlands, friend

PS1DV 11, Adele, schoolgirl, Central Midlands, daughter's friend

KBJPSUNK Unknown speaker, other

KBJPSUGP Group of unknown speakers, other

[KBP] 27179 words from *15 conversations recorded by 'Clarence' (PS065, R 511) between 13 and 19 March 1992 with 4 interlocutors*

PS065 65, Clarence, retired, Lancashire, self

PS066 67, Nina, retired, Lancashire, wife

PS067 72, Nev, retired, North-east Midlands, brother-in-law

PS068 70, Lil, retired, Lancashire, sister-in-law

KBPPSUNK Unknown speaker, other

KBPPSUGP Group of unknown speakers, other

[KBW] 115332 words from *62 conversations recorded by 'Dorothy' (PS087, R 3) between 13 and 20 March 1992 with 25 interlocutors*

PS087 34, Dorothy, teacher (pt), North-west Midlands, self

PS088 3, Tim, student (private pre), North-east Midlands, son

PS089 5, Christopher, student (state primary), North-east Midlands, son

PS08A 36, Andrew, purchasing manager, North-east Midlands, husband

PS08B 14, Elinor, student, North-east Midlands, babysitter

PS08C 40, Jane, housewife, North-east Midlands, friend

PS08D 38, Diane, housewife, North-east Midlands, friend

PS08P 40, Sue, teacher, North-east Midlands, friend

PS08S 60+, Margaret, retired, North-east Midlands, stranger

PS08T 45, Roger, teacher, friend

PS08U 60, June, teacher, Canadian, acquaintance

PS08V 60, Jean, retired, North-east Midlands, friend

PS08W 40, None, secretary, North-east Midlands, acquaintance

PS0XJ 24, Jane, teacher, North-east Midlands, friend

PS0XK 30, Dawn, housewife, North-east Midlands, friend

PS0XL 30+, Yvonne, housewife, North-east Midlands, friend

PS0XM 30+, Pat, teacher, friend

PS0XN 50, Gwen, teacher/nursery owner, North-east Midlands, friend

PS0XP 40, Lily, nursery assistant, North-east Midlands, acquaintance

PS0XR 70, Marjorie, retired, North-west Midlands, mother

PS0XS 35, Ann, teacher, son's teacher

PS0XU 50, Rosemary, housewife, Home Counties, friend

PS0XV 45, Mary, drawing plans, North-east Midlands, friend

PS0Y4 None, teacher, Home Counties, son's teacher

KBWPSUNK Unknown speaker, other

KBWPSUGP Group of unknown speakers, other

[KCC] 5311 words from *2 conversations recorded by 'Hazel' (PS0F5, R 23) on 11 January 1992 with 2 interlocutors*

PS0F5 39, Hazel, vdu clerk (pt), Welsh, self

PS0F6 7, Cerys, schoolgirl (state primary), Welsh, daughter

PS0F7 66, Marie, housewife, Welsh, mother-in-law

KCCPSUNK Unknown speaker, other

KCCPSUGP Group of unknown speakers, other

[KCF] 21898 words from *30 conversations recorded by 'Herbert' (PS1EM, R 521) between 13 and 19 March 1992 with 9 interlocutors*

PS1EM 56, Herbert, landscape gardener, Welsh, husband

PS1EN 54, Iris, home care assistant, Welsh, self

PS1EP 21, Alison, landscape gardener, Welsh, daughter

PS1ER 22, Stuart, hairdresser, Welsh, daughter's boyfriend

PS1ES 72, Gordon, retired, Welsh, friend

PS1ET 76, Mary, housewife, Welsh, mother

PS1EU 40, Mike, driver, Welsh, friend

PS1EV 80, Sally, retired, Welsh, friend

PS1EW 72, Eddie, retired, Welsh, friend

KCFPSUNK Unknown speaker, other

KCFPSUGP Group of unknown speakers, other

[KCU] 49751 words from *9 conversations recorded by 'Julie' (PS0GF, R 114) between 20 and 22 February 1992 with 6 interlocutors*

PS0GF 24, Julie, housewife, Lower South-west England, self

PS0GG 25, Gary, storesperson, Lower South-west England, husband

PS0GH 2, Vicki, pre-school, Lower South-west England, daughter

PS0GJ 20, Shelly, housewife, Lower South-west England, cousin-in-law's wife

PS0GK 25, Phil, cold store worker, Lower South-west England, cousin-in-law

PS0GL 4, Sammy, pre-school, Lower South-west England, cousin-in-law's son

KCUPSUNK Unknown speaker, other

KCUPSUGP Group of unknown speakers, other

[KCV] 32714 words from *50 conversations recorded by 'Katherine' (PS0H7, R 444) between 2 and 5 June 1991 with 3 interlocutors*

PS0H7 57, Katherine, housewife, European (German), self

PS125 56, Patrick, company secretary, husband

PS126 25, Stefan, student, European (German), friend

KCVPSUNK Unknown speaker, other

KCVPSUGP Group of unknown speakers, other

[KD0] 77692 words from *106 conversations recorded by 'Kevin' (PS0HM, R 26) between 29 November and 5 December 1991 with 14 interlocutors*

PS0HM 41, Kevin, draughtsman, London, self

PS0HN 12, Paul, student (state secondary), London, son

PS0HP 40, Ruth, teacher, wife

PS0HR 29, Michelle, local government officer, European (French), sister-in-law

PS0HS 74, Eric, retired, London, father

PS0HT 40, Adrian, salesman, London, brother-in-law

PS0HU 41, Karen, secretary, London, sister

PS0HV 33, Andrew, local government officer, London, brother

PS0HW 13, Lisa, student, London, niece

PS0HX 70+, babs (aka mutty), retired, Lower South-west England, brother-in-law's mother

PS0HY 70+, Joy, retired, London, mother

PS0J0 15, Michael, student, London, nephew

PS1KN None, ?

KD0PSUNK Unknown speaker, other

KD0PSUGP Group of unknown speakers, other

[KD1] 40487 words from *14 conversations recorded by 'Larna' (PS0JA, R 107) between 20 and 24 February 1992 with 8 interlocutors*

PS0JA 25, Larna, housewife, self

PS0JB 4, Charlotte, student (state pre), daughter

PS0JC 29, Pauline, housewife, sister-in-law

PS0JD Vicky, ?

PS0JE 25, Alex, secretary, friend

PS0JF 3, Anthony, pre-school, son

PS0JG 45, David, lorry driver, stepfather

PS0JH 3, Aaron, nephew

KD1PSUNK Unknown speaker, other

KD1PSUGP Group of unknown speakers, other

[KD3] 33516 words from *40 conversations recorded by 'Lisa' (PS0JJ, R 502) between 12 and 20 March 1992 with 7 interlocutors*

PS0JJ 30, Lisa, housewife, South Midlands, self

PS0JK 0, Peter, pre-school, son

PS0JL 29, Melvin, self-employed panel beater, husband

PS0JP 27, Diane, housewife, South Midlands, sister-in-law

PS0JR 52, Anne, maggot farm manager, South Midlands, aunt

PS0JS 27, Cheryl, housewife, neighbour

PS0JT 2, Ben, pre-school, neighbour's son

KD3PSUNK Unknown speaker, other

KD3PSUGP Group of unknown speakers, other

[KD7] 19206 words from *24 conversations recorded by 'Matt' (PS0KP, R 109) between 20 February and ?? March 1992 with 6 interlocutors*

PS0KP 35, Matt, financial advisor, Scottish, self

PS0KR 35, Jan, building society branch assistant, wife

PS0KS 13, Matthew, student (state secondary), son

PS0KT 10, Laura, student (state primary), daughter

PS0KU 7, Christopher, student (state primary), son

PS0KV Christopher, ?

KD7PSUNK Unknown speaker, other

KD7PSUGP Group of unknown speakers, other

[KD8] 76445 words from *31 conversations recorded by 'Martine' (PS0LK, R 524) between 12 and 20 March 1992 with 10 interlocutors*

PS0LK 25, Martine, senior technician, Welsh, self

PS0LL 28, Mike, construction worker, Welsh, husband

PS0LM 55, Merielle, housewife, Welsh, mother

PS0LN 45, None, pub landlord, Home Counties, stranger

PS0LP 58, Harold, engineer, Welsh, father

PS0LR 76, Nora, housewife, Welsh, grandmother-in-law

PS0LS 45, Will, civil engineer, Merseyside, colleague

PS0LT 40, Michael, technical director, Home Counties, colleague

PS0LU 27, Jim, technician, Home Counties, colleague

KD8PSUNK Unknown speaker, other

KD8PSUGP Group of unknown speakers, other

[KDD] 2206 words from *13 conversations recorded by 'Murray' (PS0KW, R 121) on 21 February 1992 with 1 interlocutor*

PS0KW 74, Murray, retired (machine operator), Canada, self

PS0KX 70, Vera, retired, wife

[KDF] 1359 words from *8 conversations recorded by 'Pamela' (PS0M6, R 65) between 31 January and 4 February 1992 with 4 interlocutors*

PS0M6 32, Pamela, sales assistant, Lancashire, self

PS0M7 58, Margaret, market researcher, Scottish, stranger

PS0M8 1, Marcus, pre-school, Lancashire, son

PS0M9 38, Paul, bus driver (psv), Lancashire, husband

KDFPSUNK Unknown speaker, other

KDFPSUGP Group of unknown speakers, other

[KDJ] 16234 words from *6 conversations recorded by 'Pauline' (PS0N3, R 117) between 21 and 24 February 1992 with 8 interlocutors*

PS0N3 26, Pauline, unemployed, North-east England, self

PS0N4 53, Bob, sales assistant, Scottish, friend

PS0N5 26, Tracey, telephonist, Scottish, friend

PS0N6 2, David, pre-school, friend

PS0N7 55, Kathy, canvasser, South Midlands, friend

PS0N8 24, Kay, canvasser, North-east Midlands, friend

PS0N9 55, Bill, canvasser, Southern (?), friend

PS0NA Michelle, North-east Midlands, ?

KDJPSUNK Unknown speaker, other

KDJPSUGP Group of unknown speakers, other

[KE2] 77961 words from *153 conversations recorded by 'Terence' (PS0W2, R 113) between 20 and 27 February 1992 with 10 interlocutors*

PS0W2 70, Terence, retired (headteacher), East Anglia, self

PS0W3 44, Richard, fireman, Lower South-west England, son

PS0W4 70, Margaret, retired, Irish, wife

PS0W5 13, Lucy, student, Lower South-west England, friend

PS0W6 13, Holly, student, Lower South-west England, friend

PS0W7 13, Adrian, student, Lower South-west England, friend

PS0W8 13, Danielle, student, Lower South-west England, friend

PS0W9 40, Christine, housewife, Lower South-west England, friend

PS0WA 50, Mima, housewife, Lower South-west England, friend

KE2PSUNK Unknown speaker, other

KE2PSUGP Group of unknown speakers, other

[KE4] 15170 words from *22 conversations recorded by 'Valerie' (PS0WN, R 75) between 30 January and ?? ?? 1992 with 7 interlocutors*

PS0WN 36, Valerie, staff nurse (pt), Scottish, self

PS0WP 34, Peter, sales representative, Scottish, husband

PS0WR 8, Jackie, student (state primary), Scottish, daughter

PS0WS 10, David, student (state primary), Scottish, son

PS0WT 11, Dawn, student, Scottish, friend

PS0WU 40+, None, dentist, Scottish, stranger

PS0WW 50+, None, telephone engineer, Scottish, stranger

PS0WX 37, Dougie, sales representative, Scottish, friend

KE4PSUNK Unknown speaker, other

KE4PSUGP Group of unknown speakers, other

[KNR] 2979 words from *3 conversations recorded by '206' (PS4XN, R 206) [dates unknown] with 5 interlocutors*

PS4XP 19, Rachel, student, Merseyside, friend

PS4XR 21, Jill, student, East Anglia, friend

PS4XS 19, Sarah, student, Lower South-west England, self

PS4XT 23, Lee, student, Lower South-west England, boyfriend

KNRPSUNK Unknown speaker, other

KNRPSUGP Group of unknown speakers, other

[KP2] 8570 words from *11 conversations recorded by 'Carla' (PS513, R 713) [dates unknown] with 8 interlocutors*

PS513 Carla, student, self

PS514 13, Papya, student, friend

PS515 Gabriel, teacher, teacher

PS516 13, Flit, student, friend

PS517 Jones, teacher, teacher

PS518 13, Marsha, student, friend

PS519 13, Sarah, student, friend

KP2PSUNK Unknown speaker, other

KP2PSUGP Group of unknown speakers, other

[KP5] 21997 words from *4 conversations recorded by 'Catherine' (PS527, R 202) on 16 October 1993 with 3 interlocutors*

PS527 20, Catherine, student, Home Counties, self

PS6P5 20, Rachel, student, Home Counties, friend

PS6TP None, ?

KP5PSUNK Unknown speaker, other

KP5PSUGP Group of unknown speakers, other

[KP7] 1806 words from *4 conversations recorded by 'Chris3' (PS52N, R 600) on 3 April 1992 with 9 interlocutors*

PS52N 35, Chris, builder, Northern England, self

PS52P 40, Rob, builder, North-east England, colleague

PS52R 45, Scon, builder, North-east England, colleague

PS52S 22, Mick, builder, North-east England, colleague

PS6P8 27, Louise, computer operator, Northern England, wife

PS6P9 36, Rosie, child minder, North-east England, friend

PS6PA 31, Phil, mortgage financier, North-east England, brother

PS6PB 2, Kyle, pre-school, North-east England, son

KP7PSUNK Unknown speaker, other

KP7PSUGP Group of unknown speakers, other

[KPU] 16953 words from *4 conversations recorded by 'Rachel2' (PS582, R 205) on 9 October 1993 with 4 interlocutors*

PS582 27, Rachel, student, Home Counties, self

PS583 29, Anne-Marie, secretary, Home Counties, sister

PS584 36, Michael, p.a., Irish, friend

PS585 40, Gearoid, p.a., Irish, friend

KPUPSUNK Unknown speaker, other

KPUPSUGP Group of unknown speakers, other

[KPX] 5165 words from *7 conversations recorded by 'Robin' (PS58K, R 723) [dates unknown] with 6 interlocutors*

PS58K 14, Robin, student, self

PS58L Jones, teacher, teacher

PS58M 13, Zoe, student, friend

PS58N 14, Ben, student, friend

PS58P 13, Oliver, student, friend

PS5AV 13, Jenny, student, friend

KPXPSUNK Unknown speaker, other

KPXPSUGP Group of unknown speakers, other

[KSN] 13743 words from *9 conversations recorded by 'Anthony2' (PS6NV, R 710) [dates unknown] with 4 interlocutors*

PS6NV 13, Anthony, student, self

PS6NW 13, Kevin, student, friend

PS6NX 10, Patrick, student, brother

KSNPSUNK Unknown speaker, other

KSNPSUGP Group of unknown speakers, other

[KSW] 6020 words from *6 conversations recorded by 'Richard4' (PS6SG, R 707) [dates unknown] with 5 interlocutors*

PS6SG 13, Richard, student, self

PS6SH 13, Frances, student, neighbour

PS6SJ 13, Lisa, student, neighbour

PS6SK Ben, friend

PS6SL Billy, friend

KSWPSUNK Unknown speaker, other

KSWPSUGP Group of unknown speakers, other

[K5K] 941 words from *The Scotsman: Applied Science pages.* u.p.

10 References

[1] Aston, G. and Burnard, L. (1998), *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

[2] Atkins, B.T.S., Clear, J., and Ostler, N. (1992), 'Corpus Design Criteria', *Literary and Linguistic Computing,* 7, pp. 1-16.

[3] Beale, A.D. (1987) 'Towards a distributional lexicon' in Garside et al (1987).

[4] Berglund, Y. (1999) 'Exploiting a large spoken corpus: An end-user's way to the BNC'. *International Journal of Corpus Linguistics* 4:29-52.

[5] Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan (1999) *The Longman grammar* Harlow: Pearson Education Limited

[6] Brill, E. (1992) 'A simple rule-based part-of-speech tagger' in *Proceedings of the 3rd conference on Applied Natural Language Processing.* Italy: Trento.

[7] Burnage, G. and Dunlop, D. (1992) 'Encoding the British National Corpus' in Aarts et al, eds. *English language corpora: design, analysis and exploitation* Amsterdam: Rodopi, pp 79-95

[8] Burnard, L. (1999) 'Using SGML for linguistic analysis: the case of the BNC' in *Markup languages theory and practice*. I.2 pp. 31-51. Cambridge, Mass: MIT Press. Also published in *Maschinelle Verarbeitung altdeutscher Texte V*, pp 53-72. Tuebingen: Max Niemeyer, 2001.

[9] Burnard, L. (2002) 'Where did we go wrong? a retrospective look at the British National Corpus' in *Teaching and learning by doing corpus analysis*, ed. B. Kettemann and G. Markus. Amsterdam: Rodopi, pp 51-71.

[10] Clear, J. H. (1993) 'The British National Corpus' in Delany, P. and Landow, G., ed. *The Digital Word : text-based computing in the humanities*. Cambridge (Mass), MIT Press, pp. 163-187.

[11] Crowdy, S. (1994) 'Spoken Corpus Transcription' in *Literary and Linguistic Computing*, 9.1 pp 25-28.

[12] Crowdy, S. (1995) 'The BNC spoken corpus' in Leech, G., Myers, G. and Thomas, J., eds. *Spoken English on computer: transcription, mark-up and application* Harlow: Longman, pp. 224-235.

[13] Dunlop, D. (1995) 'Practical considerations in the use of TEI headers in large corpora' in Ide, N. and Veronis, J. eds. *Text Encoding Initiative: background and context.* Dordrecht: Kluwer, pp 85-98.

[14] Fligelstone, S., Rayson, P., and Smith, N. (1996). 'Template analysis: bridging the gap between grammar and the lexicon' in Thomas, J. and Short, M., eds. *Using corpora for language research* Harlow: Longman, pp 181-207.

[15] Fligelstone S., Pacey M., and Rayson P. (1997) 'How to Generalize the Task of Annotation'. In Garside et al. (1997)

[16] Garside R., Leech G. and Sampson, G. (eds.) (1987) *The Computational Analysis of English.* London: Longman.

[17] Garside R., Leech G. and McEnery A. (eds.) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

[18] Garside R., and Smith N. (1997) 'A hybrid grammatical tagger: CLAWS4'. In Garside et al. (1997)

[19] Lehmann, H., Schneider, P., Hoffmann, S. (1999) 'BNCweb' in Kirk, J. ed. *Corpora galore: analysis and techniques in describing English.* Amsterdam: Rodopi, pp. 259-266.

[20] Lee, D. (2001) ‘Genres, registers, text types and styles: clarifying the concepts and navigating a path through the BNC Jungle’ in *Language Learning and Technology*, vol 5 no 3, September 2001; available online at `http://llt.msu.edu/vol5num3/lee/default.html`

[21] Leech, G., Garside, R., and Bryant, M. (1994). ‘CLAWS4: The tagging of the British National Corpus’ in *Proceedings of the 15th International Conference on Computational Linguistics (COLING 94)*. Japan: Kyoto. (pp.622-628.)

[22] Leech, G., Rayson, P., and Wilson, A. (2001). *Word frequencies in written and spoken English based on the British National Corpus*. London: Longman. Extracts available online at `http://www.comp.lancs.ac.uk/ucrel/bncfreq/`

[23] Marshall, I. (1983). ‘Choice of Grammatical wordclass without Global Syntactic Analysis: Tagging Words in the LOB Corpus’ in *Computers and the Humanities* 17, 139-50.

[24] Smith, N. (1997) ‘Improving a Tagger’ in Garside et al. (1997)

[25] Sperberg-McQueen, C.M. and Burnard, L (eds) (1994). *Guidelines for Electronic Text Encoding and Interchange (TEI P3)*. Oxford, Virginia, Brown: Text Encoding Initiative.

[26] Sperberg-McQueen, C.M. and Burnard, L (eds) (2007). *TEI Guidelines, P5*. Oxford, Virginia, Brown: Text Encoding Initiative; available online at `http://www.tei-c.org/P5/`.