What is the BNC?

The British National Corpus (BNC) is:

- a 100 million word collection of
- samples of written and spoken language from
- a wide range of sources,
- designed to represent a wide cross-section of
- British English from
- the later part of the 20th century, both
- spoken and written.

The latest edition is the *BNC XML Edition*, released in 2007.

What is the BNC?

The **written part** of the BNC (90%) includes [...]

- extracts from regional and national newspapers,
- specialist periodicals and journals for all ages and interests,
- academic books [...]
- popular fiction,
- published and unpublished letters and memoranda,
- school and university essays, [...]
- many other kinds of text.

What is the BNC?

The **spoken part** (10%) consists of orthographic transcriptions of

- unscripted informal conversations (recorded by volunteers selected from
  - different age, region and social classes
  - in a demographically balanced way) and
- spoken language collected in different contexts, ranging from
  - formal business or government meetings to
  - radio shows and
  - phone-ins.

What is the BNC?

The corpus is **encoded** according to the Guidelines of the Text Encoding Initiative (TEI) to represent both

- the output from CLAWS (automatic part-of-speech tagger) and
- a variety of other structural properties of texts (e.g.
  - headings,
  - paragraphs,
  - lists etc.).
- Full classification, contextual and bibliographic information is also included with each text in the form of a TEI-conformant header.

What is the BNC?

Work on building the corpus began in 1991, and was completed in 1994.

No new texts have been added after the completion of the project but the corpus was slightly revised prior to the release of the second edition *BNC World* (2001) and the third edition *BNC XML Edition* (2007).

Since the completion of the project, two sub-corpora with material from the BNC have been released separately:

- the BNC Sampler (a general collection of one million written words, one million spoken) and
- the BNC Baby (four one-million word samples from four different genres).

What sort of corpus is the BNC?

- **Monolingual:** It deals with modern British English, not other languages used in Britain. However non-British English and foreign language words do occur in the corpus.

- **Synchronic:** It covers British English of the late twentieth century, rather than the historical development which produced it.

- **General:** It includes many different styles and varieties, and is not limited to any particular subject field, genre or register. In particular, it contains examples of both spoken and written language.

- **Sample:** For written sources, samples of 45,000 words are taken from various parts of single-author texts. Shorter texts up to a maximum of 45,000 words, or multi-author texts such as magazines and newspapers, are included in full. Sampling allows for a wider coverage of texts within the 100 million limit, and avoids over-representing idiosyncratic texts.

The BNC consortium

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
- 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.

Maintenance and distribution of the BNC

Maintenance, distribution, and development of the corpus has been carried out at Oxford University Computing Services.

There have been three major revisions of the corpus:

- BNC 1.0 (1995)
- BNC World Edition (2000)
- BNC XML Edition (2007)

BNC User Reference Guide
(http://www.natcorp.ox.ac.uk/XMLedition/URG/intro.html)

- **Section 1 Design of the corpus describes the planned uses of BNC and the design of the spoken and the written components**

- **Section 2 Basic structure describes the basic structure of the BNC encoding scheme, in terms of the XML elements and attributes distinguished and the tags used to mark them.**

- **Section 3 Written texts describes features which are peculiar to written texts, and**

- **[S]ection 4 Spoken texts those peculiar to spoken texts.**

BNC User Reference Guide
(http://www.natcorp.ox.ac.uk/XMLedition/URG/intro.html)

- In each case, a distinction is made between those elements which are marked up in all texts and those which (for technical or financial reasons) are not always so distinguished, and hence appear in some texts only. It should be noted that by no means all of the features described here will be present in every text of the corpus, nor, if present, will they necessarily be tagged.

- Section 5 The header describes the structure of the detailed metadata associated with each text, in the form of the <teiHeader> element attached to each component of the corpus, and also to the whole corpus itself.

- Sections 6 through 12 covers the linguistic annotation, the software and the sources

BNC User Reference Guide: 1 Design of the corpus

(http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

The uses originally envisaged for the British National Corpus were set out in a working document called *Planned Uses of the British National Corpus* BNCW02 (11 April 91). This document identified the following as likely application areas for the corpus:

- **reference book publishing**
- **academic linguistic research**
- **language teaching**
- **artificial intelligence**
- **natural language processing**
- **speech processing**
- **information retrieval**

BNC User Reference Guide: 1 Design of the corpus (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

The same document identified the following categories of linguistic information derivable from the corpus:

- lexical
- semantic/pragmatic
- syntactic
- morphological
- graphological/written form/orthographical

In the 15 or more years since that document was published, it has become apparent that the corpus, and corpus methods in general, have had a far wider impact than anticipated, notably in the field of language teaching.

BNC User Reference Guide: 1 Design of the corpus. Composition. (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

There is a broad consensus among the participants in the project and among corpus linguists that a general-purpose corpus of the English language would ideally contain a high proportion of spoken language in relation to written texts. However, it is significantly more expensive to record and transcribe natural speech than to acquire written text in computer-readable form. Consequently the spoken component of the BNC constitutes approximately 10 per cent (10 million words) of the total and the written component 90 per cent (90 million words). These were agreed to be realistic targets, given the constraints of time and budget, yet large enough to yield valuable empirical statistical data about spoken English. In the BNC sampler, a two per cent sample taken from the whole of the BNC, spoken and written language are present in approximately equal proportions, but other criteria are not equally balanced.

BNC User Reference Guide: 1 Design of the corpus. Composition (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

From the start, a decision was taken to select material for inclusion in the corpus according to an overt methodology, with specific target quantities of clearly defined types of language. This approach makes it possible for other researchers and corpus compilers to review, emulate or adapt concrete design goals. This section outlines these design considerations, and reports on the final make-up of the BNC.

This and the other tables in this section show the actual make-up of the second version of the British National Corpus (the BNC World Edition) in terms of

- texts : number of distinct samples not exceeding 45,000 words
- S-units: number of <s> elements identified by the CLAWS system (more or less equivalent to sentences)
- W-units: number of <w> elements identified by the CLAWS system (more or less equivalent to words)

BNC User Reference Guide: 1 Design of the corpus. Composition (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

The XML Edition of the BNC contains 4049 texts and occupies (including all markup) 5,228,040 Kb, or about 5.2 Gb. In total, it comprises just under 100 million orthographic words (specifically, 96986707), but the number of w-units (POS-tagged items) is slightly higher at 98363783. The tagging distinguishes a further 13614425 punctuation strings, giving a total content count of 110691482 strings. The total number of s-units tagged is about 6 million (6026284). Counts for these and all the other elements tagged in the corpus are provided in the corpus header.

In the following tables both an absolute count and a percentage are given for all the counts. The percentage is calculated with reference to the relevant portion of the corpus, for example, in the table for "written text domain", with reference to the total number of w-units in written texts. Note that punctuation strings are not included in these totals. The reference totals used are given in the first table below.

BNC User Reference Guide: 1 Design of the corpus. Composition

(http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Table 1. Text type

	texts	w-units	%	s-units	%
Spoken demographic	153	4233955	4.30	610557	10.13
Spoken context-governed	755	6175896	6.27	427523	7.09
Written books and periodicals	2685	79238146	80.55	4395581	72.94
Written-to-be-spoken	35	1278618	1.29	104665	1.73
Written miscellaneous	421	7437168	7.56	487958	8.09

BNC User Reference Guide: 1 Design of the corpus. Composition (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

All texts are also classified according to their date of production. For spoken texts, the date was that of the recording. For written texts, the date used for classification was the date of publication of the source edition used, for the most part; in the case of imaginative works, however, the date of first publication of the work was used. Informative texts were selected only from 1975 onwards, imaginative ones from 1960, reflecting their longer ‘shelf-life’, though most (75 per cent) of the latter were published no earlier than 1975.

Table 2. Publication date

	texts	w-units	%	s-units	%
Unknown	162	1831585	1.86	126416	2.09
1960-1974	46	1718449	1.74	119510	1.98
1975-1984	169	4730889	4.80	257962	4.28
1985-1993	3672	90082860	91.58	5522396	91.63

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Sampling basis: production and reception

While it is sometimes useful to distinguish in theory between language which is received (read and heard) and that which is produced (written and spoken), it was agreed that the selection of samples for a general-purpose corpus must take account of both perspectives.

Text that is published in the form of books, magazines, etc., is not representative of the totality of written language that is produced, as writing for publication is a comparatively specialized activity in which few people engage.

BNC User Reference Guide: 1 Design of the corpus. **1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)**

However, it is much more representative of written language that is received, and is also easier to obtain in useful quantities, and thus forms the greater part of the written component of the corpus.

There was no single source of information about published material that could provide a satisfactory basis for a sampling frame, but a combination of various sources furnished useful information about the totality of written text produced and, particularly, received, some sources being more significant than others.

They are principally statistics about books and periodicals that are published, bought or borrowed.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Catalogues of **books published per annum** tell us something about production but little about reception as many books are published but hardly read.

A list of **books in print** provides somewhat more information about reception as time will weed out the books that nobody bought (or read): such a list will contain a higher proportion of books that have continued to find a readership.

The books that have the widest reception are presumably those that figure in **bestseller lists**, particularly **prize winners** of competitions such as the Booker or Whitbread.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Such works were certainly candidates for inclusion in the corpus, but the statistics of book-buying are such that very few texts achieve high sales while a vast number sell only a few or in modest numbers.

If texts had been selected in strict arithmetical proportion to their sales, their range would have been severely limited. However, where a text from one particular subject domain was required, it was appropriate to prefer a book which had achieved high sales to one which had not.

Library lending statistics, where these are available, also indicate which books enjoy a wide reception and, like lists of books in print, show which books *continue* to be read.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Similar observations hold for magazines and periodicals. lists of **current magazines and periodicals** are similar to catalogues of published books, but perhaps more informative about language reception, as it may be that periodicals are bought and read by a wider cross-section of the community than books. Also, a periodical that fails to find a readership will not continue to be published for long.

Periodical circulation figures have to be treated with the same caution as bestseller lists, as a few titles dominate the market with a very high circulation. To concentrate too exclusively on these would reduce the range of text types in the corpus and make contrastive analysis difficult.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Published written texts were selected partly at random from Whitaker's *Books in Print* for 1992 and partly systematically, according to the selection features outlined in section Selection features below.

Available sources are concerned almost exclusively with published books and periodicals. It is much more difficult to obtain data concerning the production or reception of unpublished writing. Intuitive estimates were therefore made in order to establish some guidelines for text sampling in the latter area.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Selection features

Texts were chosen for inclusion according to three selection features:

- domain (subject field),
- time (within certain dates) and
- medium (book, periodical, etc.).

The purpose of these selection features was to ensure that the corpus contained a broad range of different language styles, for two reasons. The first was so that the corpus could be regarded as a microcosm of current British English in its entirety, not just of particular types. The second was so that different types of text could be compared and contrasted with each other.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Selection Procedure

Each selection feature was divided into classes (e.g. ‘Medium’ into books, periodicals, unpublished etc.; ‘Domain’ into imaginative, informative, etc.) and target percentages were set for each class.

These percentages are quite independent of each other: there was no attempt, for example, to make 25 per cent of the selected periodicals imaginative.

The design proposed that seventy-five per cent of the samples be drawn from informative texts, and the remaining 25 per cent from imaginative texts.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

It further proposed that titles be taken from a variety of media, in the following proportions:

- 60 per cent from books,
- 30 per cent from periodicals,
- 10 per cent from miscellaneous sources (published, unpublished, and written to be spoken).

Half of the books in the ‘Books and Periodicals’ class were selected at random from Whitaker's *Books in Print 1992*. This was to provide a control group to validate the categories used in the other method of selection: the random selection disregarded Domain and Time, but texts selected by this method were classified according to these other features after selection.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Sample size and method
For books, a target sample size of 40,000 words was chosen.

No extract included in the corpus exceeds 45,000 words.

For the most part, texts which in their entirety were shorter than 40,000 words were further reduced by ten per cent for copyright reasons; a few texts longer than the target size were however included in their entirety.

Text samples normally consist of a continuous stretch of discourse from within the whole. A convenient breakpoint (e.g. the end of a section or chapter) was chosen as far as possible to begin and end the sample so that high-level discourse units were not fragmented.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Where possible, no more than one sample was taken from any one text; for newspaper texts and large encyclopaedic works, no sample greater than 40,000 words was taken.

Samples were taken randomly from the beginning, middle or end of longer texts. (In cases where a publication included essays or articles by a variety of authors of different nationalities, the work of non-UK authors was omitted.)

Some types of written material are composite in structure: that is, the physical object in written form is composed of more than one text unit. Important examples are issues of a newspaper or magazine which, though editorially shaped as a document, contain discrete texts, each with its specific authorship, stylistic characteristics, register and domain.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

The BNC attempts to separate these discrete texts where appropriate and to classify them individually according to the selection and classification features. As far as possible, the individual stories in one issue of a newspaper were grouped according to domain, for example as ‘Business’ articles, ‘Leisure’ articles, etc.

The following subsections discuss each selection criterion, and indicate the actual numbers of words in each category included.

BNC User Reference Guide: 1 Design of the corpus. **1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)**

Domain

Classification according to subject field seems hardly appropriate to texts which are fictional or which are generally perceived to be literary or creative.

Consequently, these texts are all labelled imaginative and are not assigned to particular subject areas.

All other texts are treated as informative and are assigned to one of the eight domains listed below.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Table 3. Written Domain

	texts	w-units	%	s-units	%
Imaginative	476	16496420	18.75	1352150	27.10
Informative: natural & pure science	146	3821902	4.34	183384	3.67
Informative: applied science	370	7174152	8.15	356662	7.15
Informative: social science	526	14025537	15.94	698218	13.99
Informative: world affairs	483	17244534	19.60	798503	16.00
Informative: commerce & finance	295	7341163	8.34	382374	7.66
Informative: arts	261	6574857	7.47	321140	6.43
Informative: belief & thought	146	3037533	3.45	151283	3.03
Informative: leisure	438	12237834	13.91	744490	14.92

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

The evidence from catalogues of books and periodicals suggests that imaginative texts account for significantly less than 25 per cent of published output, and unpublished reports, correspondence, reference works and so on would seem to add further to the bulk of informative text which is produced and consumed. However, the overall distribution between informative and imaginative text samples is set to reflect the influential cultural role of literature and creative writing. The target percentages for the eight informative domains were arrived at by consensus within the project, based loosely upon the pattern of book publishing in the UK during the past 20 years or so, as reflected in the categorized figures for new publications that appear annually in Whitaker's *Book list*.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component

(http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Medium

This categorisation is broad, since a detailed taxonomy or feature classification of text medium could have led to such a proliferation of subcategories as to make it impossible for the BNC adequately to represent all of them.

The labels used here are intended to be comprehensive in the sense that any text can be assigned with reasonable confidence to these macro categories.

The labels we have adopted represent the highest levels of a fuller taxonomy of text medium.

BNC User Reference Guide: 1 Design of the corpus. **1.4 Design of the written component** **(http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)**

Table 4. Written Medium

	texts	w-units	%	s-units	%
Book	1411	50293803	57.18	2887523	57.88
Periodical	1208	28609494	32.52	1487644	29.82
Miscellaneous published	238	4233135	4.81	287700	5.76
Miscellaneous unpublished	249	3538882	4.02	220672	4.42
To-be-spoken	35	1278618	1.45	104665	2.09

The ‘Miscellaneous published’ category includes brochures, leaflets, manuals, advertisements. The ‘Miscellaneous unpublished’ category includes letters, memos, reports, minutes, essays. The ‘written-to-be-spoken’ category includes scripted television material, play scripts etc.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Descriptive features
Written texts may be further classified according to sets of descriptive features.

These features *describe* the sample texts; they did not determine their selection.

This information is recorded to allow more delicate contrastive analysis of particular sets of texts.

As a simple example, the gross division into two time periods in the selection features can, of course, be refined and subcorpora defined over the BNC for more specific dates.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

However, the relative sizes of such subcorpora are undefined by the BNC design specification.

These descriptive features were monitored during the course of the data gathering, and text selection, in cases where a free choice of texts was available, took account of the relative balance of these features.

Thus although no relative proportions were defined for different target age groups (for example), we ensured that the corpus does contain texts intended for children as well as for adults.

The following tables summarize the results for the first release of the corpus. Note that many texts remain unclassified.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Author information
Information about authors of written texts was included only where it was readily available, for example from the dust-wrapper of a book. Consequently, the coverage of such information is very patchy.

The authorship of a written text was characterized as ‘corporate’ where it was produced by an organization and no specific author was given, and as ‘multiple’ in cases where several authors were named.

Author sex was classified as ‘mixed’ where more than one author of either sex was specified, and ‘unknown’ where it could not reliably be determined from the author's name.

Note that ‘author age’ means the author's age at the time of creation of the work concerned.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Table 5. Author type

	texts	w-units	%	s-units	%
Unknown	211	3786835	4.30	174371	3.49
Corporate author	347	6497144	7.38	455649	9.13
Multiple author	1322	34563219	39.29	1810901	36.30
Sole author	1261	43106734	49.01	2547283	51.06

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Table 6. Sex of author

	texts	w-units	%	s-units	%
Unknown	1573	36161115	41.11	1968162	39.45
Author sex Male	920	30665582	34.86	1671420	33.50
Author sex Female	414	14588260	16.58	967522	19.39
Author sex Mixed	234	6538975	7.43	381100	7.64

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Table 7. Author age-group

	texts	w-units	%	s-units	%
Unknown	2518	66000719	75.04	3687586	73.92
Author age 0-14	3	59559	0.06	3443	0.06
Author age 15-24	19	542578	0.61	29810	0.59
Author age 25-34	66	2267139	2.57	159455	3.19
Author age 35-44	191	6726926	7.64	410143	8.22
Author age 45-59	205	7230714	8.22	410644	8.23
Author age 60+	139	5126297	5.82	287123	5.75

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Table 8. Domicile

	texts	w-units	%	s-units	%
Unknown	2272	57227155	65.06	3133068	62.80
Author domicile UK and Ireland	841	29760000	33.83	1798301	36.05
Author domicile Commonwealth	12	411207	0.46	25759	0.51
Author domicile Continental Europe	6	234402	0.26	12466	0.24
Author domicile USA	8	245604	0.27	15675	0.31
Author domicile Elsewhere	2	75564	0.08	2935	0.05

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Target audience
Some attempt was made to characterize the kind of audience for which written texts were produced in terms of

- age,
- sex and
- ‘level’ (a subjective assessment of the text's technicality or difficulty).

The last of these proved very difficult to assess and was very frequently confused with circulation size or audience size; for that reason, no figures for it are included here.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Table 9. Audience age

	texts	w-units	%	s-units	%
Child audience	42	903690	1.02	81074	1.62
Teenager audience	78	1831178	2.08	138098	2.76
Adult audience	2911	81928776	93.14	4597388	92.16
Any audience	110	3290288	3.74	171644	3.44

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Table 10. Audience sex

	texts	w-units	%	s-units	%
Unknown	706	20271270	23.04	1131254	22.67
Male audience	61	2396935	2.72	135950	2.72
Female audience	175	6904137	7.84	503629	10.09
Mixed audience	2199	58381590	66.37	3217371	64.49

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component

(http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Miscellaneous classification information: Written texts were also characterized according to their place of publication and the type of sampling used.

Table 11. Publication place

	texts	w-units	%	s-units	%
Unknown	690	14718827	16.73	788440	15.80
UK (unspecific) publication	263	7163111	8.14	380824	7.63
Ireland publication	37	570652	0.64	31793	0.63
UK: North (north of Mersey-Humber line) publication	191	3781055	4.29	228247	4.57
UK: Midlands (north of Bristol Channel-Wash line) publication	93	2590345	2.94	177308	3.55
UK: South (south of Bristol Channel-Wash line) publication	1853	58587808	66.61	3360401	67.36
United States publication	14	542134	0.61	21191	0.42

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Table 12. Sampling type

	texts	w-units	%	s-units	%
Unknown	1583	35551102	40.42	1991798	39.93
Whole text	270	6524975	7.41	433722	8.69
Beginning sample	584	21075222	23.96	1119251	22.43
Middle sample	510	18454807	20.98	1049692	21.04
End sample	119	4317326	4.90	253322	5.07
Composite sample	75	2030500	2.30	140419	2.81

In addition to the above, standard bibliographic details such as author, title, publication details, extent, topic keywords etc. were recorded for the majority of texts, as further described below (see 5 The header).

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Selection procedures employed

Books

Roughly half the titles were randomly selected from available candidates identified in Whitaker's *Books in Print* (BIP), 1992, by students of Library and Information Studies at Leeds City University. Each text randomly chosen was accepted only if it fulfilled certain criteria: it had to be published by a British publisher, contain sufficient pages of text to make its incorporation worthwhile, consist mainly of written text, fall within the designated time limits, and cost less than a set price. The students noted the ISBN, author, title and price of each book thus selected; the final selection weeded out texts by non-UK authors.

Half of the books having been selected by this method, the remaining half were selected systematically to make up the target percentages in each category. The selection proceeded as follows.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Bestsellers

Because of their wide reception, bestsellers were obvious candidates for selection. The lists used were those that appeared in the *Bookseller* at the end of the years 1987 to 1993 inclusive. Some of the books in the lists were rejected, for a variety of reasons. Obviously books that had already been selected by the random method were excluded, as were those by non-UK authors. In addition, a limit of 120,000 words from any one author was imposed, and books belonging to a domain or category whose quota had already been reached were not selected. Other bestseller lists were obtained from *The Guardian,* the British Council, and from Blackwells Paperback Shop.

The titles yielded by this search were mostly in the Imaginative category.

BNC User Reference Guide: 1 Design of the corpus. **1.4 Design of the written component**
(http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Literary prizes
The criteria for inclusion were the same as for bestsellers. The prize winners, together with runners-up and shortlisted titles, were taken from several sources, principally Anne Strachan, *Prizewinning literature: UK literary award winners,* London, 1989. For 1990 onwards the sources used were: the last issue of the *Bookseller* for each year; *The Guardian Index, 1989–*, entries under the term 'Literature'; and *The Times Index, 1989-,* entries under the term 'Literature — Awards'.
Literary prizes are in the main awarded to works that fall into the Imaginative category, but there are some Informative ones also.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component
(http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Library loans

The source of statistics in this category was the record of loans under Public Lending Right, kindly provided by Dr J. Parker, the Registrar. The information comprised lists of the hundred most issued books and the hundred most issued children's books, in both cases for the years 1987 to 1993.

The lists consist almost exclusively of imaginative literature, and many titles found there also appear in the lists of bestsellers and prize winners.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Additional texts

As collection proceeded, monitoring disclosed potential shortfalls in certain domains. A further selection was therefore made, based on the ‘Short Loan’ collections of seven University libraries. (Short Loan collections typically contain books required for academic courses, which are consequently in heavy demand.)

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Periodicals and magazines
Periodicals, magazines and newspapers account for 30 per cent of the total text in the corpus. Of these, about 250 titles were issues of newspapers. These were selected to cover as wide a spectrum of interests and language as possible. Newspapers were selected to represent as wide a geographic spread as possible: *The Scotsman* and the *Belfast Telegraph* are both represented, for example.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Other media

In addition to samples from books, periodicals, and magazines, the written part of the corpus contains about seven million words classified as ‘Miscellaneous Published’, ‘Miscellaneous Unpublished’, or as ‘Written to be spoken’. The distinction between ‘published’ and ‘unpublished’ is not an easy one; the former category largely contains publicity leaflets, brochures, fact sheets, and similar items, while the latter has a substantial proportion of school and university essays, unpublished creative writing or letters, and internal company memoranda. The ‘written to be spoken’ material includes scripted material, intended to be read aloud such as television news broadcasts; transcripts of more informal broadcast materials such as discussions or phone-ins are included in the spoken part of the corpus.

BNC User Reference Guide: 1 Design of the corpus. 1.4 Design of the written component (http://www.natcorp.ox.ac.uk/XMLedition/URG/BNCdes.html)

Copyright permissions

Before a selected text could be included, permissions had to be obtained from the copyright owner (publisher, agent, or author). A standard Permissions Request was drafted with considerable care, but some requests were refused, or simply not answered even after prompting, so that the texts concerned had to be excluded or replaced.