British National Corpus

About the British National Corpus

Contents

- What is the BNC?
  - What sort of corpus is the BNC?
- How the BNC was created
  - Creation process in brief
- The BNC in numbers
- BNC Products
  - BNC XML Edition
  - BNC Baby
  - BNC Sampler
  - Brown Corpus
  - BNC World
- SARA/XAIRA
- BNC Consortium

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.

The **written part** of the BNC (90%) includes, for example, extracts from regional and national newspapers, specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text. The **spoken part** (10%) includes a large amount of unscripted informal conversation, recorded by volunteers selected from different age, region and social classes in a demographically balanced way, together with spoken language collected in all kinds of different contexts, ranging from formal business or government meetings to radio shows and phone-ins.

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.

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).

Full technical documentation covering all aspects of the BNC including its design, markup, and contents are provided by the Reference Guide for the British National Corpus (XML Edition). For earlier versions of the Reference Guide and other documentation, see the BNC Archive page.

What sort of corpus is the BNC?

Monolingual: It deals with modern British English, not other languges 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.

How the BNC was created

The BNC project was carried out and is managed by the **BNC Consortium**, an industrial/academic consortium led by Oxford University Press, of which the other members are major dictionary publishers Addison-Wesley Longman and Larousse Kingfisher Chambers; academic research centres at Oxford University Computing Services (OUCS), the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University, and the British Library's Research and Innovation Centre. The project was funded by the commercial partners, the Science and Engineering Council (now EPSRC) and the DTI under the Joint Framework for Information Technology (JFIT) programme. Additional support was provided by the British Library and the British Academy.

The design principles and the creation of the corpus are described in more detail on a separate page:

Creation process in brief

The creation of the corpus started with a careful planning stage where the design principles were drawn up. These principles included the selection criteria that were used as the basis for the collection of the texts (a separate section describes the selection criteria for the written and the spoken parts of the corpus).

Once a suitable texts was identified and permission to use it had been obtained, the text was converted to machine readable form. The conversion was performed by one of the commercial partners (OUP, Longman or Chambers). The resulting text was then converted to the standard project encoding format at OUCS, where its accuracy and internal consistency was also validated. The text was then passed to UCREL, where word class tagging was automatically added, and returned to OUCS for documentation and accession into the corpus. Each stage of corpus processing was recorded in a database maintained at OUCS.

Work on building the corpus commenced in 1991 and was completed in 1994. The first general release of the Corpus for European researchers was announced in February 1995. 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*. BNC World was made available for world-wide distribution in 2001. In response to user feedback, the original SGML version of the corpus was later converted into XML. Additional mark-up for lemma and simplified word-class annotation was added and the treatment of multi-word units was improved. Minor errors and inconsistencies were also corrected. BNC XML Edition was released in 2007. Two sub-sets from the corpus have been published separately: the *BNC Sampler* and the *BNCBaby*.

The design principles and the creation of the corpus are described in more detail on a separate page:

The BNC in numbers

[This page is currently being revised. For BNC XML Edition figures, see the Reference Guide for the British National Corpus (XML Edition).] The BNC World Edition contains 4,054 texts and occupies (including SGML markup) 1,508,392 Kbytes, or about 1.5 Gb. In total, it comprises just over 100 million orthographic words (specifically 100,467,090), but the number of w-units (POS-tagged items) is slightly less: 97,619,934. The total number of s-units identified by CLAWS is just over 6 million (6,053,093). To put these numbers into perspective, the average paperback book has about 250 pages per centimetre of thickness; assuming 400 words a page, we calculate that the whole corpus printed in small type on thin paper would take up about ten metres of shelf space. Reading the whole corpus aloud at a fairly rapid 150 words a minute, eight hours a day, 365 days a year, would take just over four years.

Table 1. Composition of the BNC World Edition

Text type	Texts	Kbytes	W-units	S-units	percent
Spoken demographic	153	4206058	4.30	610563	10.08
Spoken context-governed	757	6135671	6.28	428558	7.07
All Spoken	910	10341729	10.58	1039121	17.78
Written books and periodicals	2688	78580018	80.49	4403803	72.75
Written-to-be-spoken	35	1324480	1.35	120153	1.98
Written miscellaneous	421	7373707	7.55	490016	8.09
All Written	3144	87278205	89.39	5013972	82.82

More BNC World frequency tables are available in the BNC User Reference Guide.

Word frequency lists have been produced and published online by, for example, Leech, Rayson and Wilson (see Word Frequencies in Written and Spoken English: based on the British National Corpus and Adam Kilgarriff (available from his website).

BNC Products

The British National Corpus (BNC) Consortium was formed in 1990, and started work in 1991 on the three-year task of producing a hundred-million word corpus of modern British English

for use in commercial and academic research. The full BNC contains about 100 million words: 90% written, 10% orthographically transcribed spoken text. The first edition was published in 1994. A slightly revised version, *BNC World*, was made available world-wide in 2001. In 2007, the BNC was made available in XML. *BNC XML Edition* is the version currently distributed and supported. Two subsets of the BNC have been produced separately: *BNC Sampler* and *BNC Baby*.

The BNC corpora are distributed with a search tool. *Xaira* is a development of the SARA program originally developed for use with the first versions of the BNC and BNC Sampler. *XAIRA* can be used with the BNC Baby, BNC Sampler (XML-version) and BNC XML Edition as well as with other corpora in XML format.

BNC XML Edition

The BNC contains about 100 million words: 90% written, 10% orthographically transcribed spoken text. It has been annotated with word-class information (part-of-speech) and the texts also contain metatextual information.*BNC XML Edition* is a revised version of the BNC World and it was released in 2007. BNC XML Edition has some additional information about lemmas and simplified word-class of the individual words, but apart from a few errors and inconsistencies, no changes have been made to the actual corpus texts between the two versions. This version of the corpus is in XML format and can be used with the XAIRA search program which allows more search options and an improved user interface than the previous SARA program.

BNC XML Edition is made available on DVD for installation on a stand-alone PC or on a Windows, Unix or OSX server. It is delivered with a copy of the XAIRA search program and all necessary XAIRA index files.

For more information about the BNC XML Edition corpus, follow the links to the Reference Guide for the British National Corpus (XML Edition). Information about the BNC project and the original creation of the corpus can be found at corpus creation page. To buy a copy of the corpus, follow the links to the How to order page.

BNC Baby

Figure 1. BNC Baby CD cover

BNC Baby is a subset of the BNC World. It consists of four one-million word samples, each compiled as an example of a particular genre: fiction, newspapers, academic writing and spoken conversation. The texts have the same annotation as the full corpus (part of speech, meta data, etc). The Reference Guide for BNC-baby offers further information about this sample, such as a description of the design and information about the way in which it is encoded.

The BNC Baby is in XML format and can be searched with the XAIRA Tool. It is distributed on a CD together with the *BNC Sampler* and an XML version of the American English *Brown corpus*. More information about the CD is available on the BNC Baby CD page. The CD can be ordered online.

BNC Sampler

The BNC Sampler is a subset of the full BNC. It comprises two samples of written and spoken material of one million words each, compiled to mirror the composition of the full BNC as far as possible. The word-class annotation of the BNC Sampler texts has been carefully checked and manually corrected. The Sampler was first created at Lancaster University during the creation of the BNC.

The BNC Sampler is in XML format and can be searched with the XAIRA Tool. It is distributed on the BNC Baby CD together with the *BNC Baby* and an XML version of the American English *Brown corpus*. How to order

Brown Corpus

The Brown Corpus of Standard American English was created at the Brown university by by W. N. Francis and H. Kucera. It contains one million words of written American English, taken from publications from 1961. The texts are all appr, 2,000 words long and grouped into 15 categories. More information about the content of the corpus can be found in the *Brown Corpus Manual* by Francis and Kucera, available on the ICAME webpage.

This version of the Brown corpus has word-class annotation and has been converted into XML and indexed to be used with the XAIRA program. It is distributed on the BNC Baby CD together with the *BNC Baby* and *BNC Sampler* corpora. How to order

BNC World

The BNC contains about 100 million words: 90% written, 10% orthographically transcribed spoken text. *BNC World* is a revised version of the original BNC and was produced between 1998 and 2000. It contains a thorough revision of the part of speech tagging, several corrections to the headers, and some minor revision of the SGML tagging used. BNC World was made available world-wide in 2001. It has now been superseded by BNC XML Edition.

BNC World was made available on CD for installation on a stand-alone PC or on a Windows, Unix or OSX server. The corpus can also be accessed via the BNC Subscription service or by using the BNC Simple Search.

For more information about the BNC World corpus, follow the links to the Users Reference Guide.

SARA/XAIRA

All versions and subsets of the BNC are delivered with a search program: SARA or XAIRA (depending on the format of the corpus). SARA (SGML-Aware Search Application) can be used with BNC World (or other corpora in SGML). XAIRA (XML Aware Indexing and Retrieval Architecture) is the tool distributed on the BNC Baby CD (for use with BNC Baby, BNC Sampler or Brown corproa) and with the BNC XML Edition. XAIRA is an open-source application that can also be used with other corpora or texts in XML-format. For more information about SARA or XAIRA, visit the SARA/XAIRA page.

BNC Consortium

The BNC project was carried out and is managed by the **BNC Consortium**, an

industrial/academic consortium lead by Oxford University Press, of which the other members are major dictionary publishers Addison-Wesley Longman and Larousse Kingfisher Chambers; academic research centres at Oxford University Computing Services (OUCS), the University Centre for Computer Corpus Research on Language (UCREL) at Lancaster University, and the British Library's Research and Innovation Centre. The Consortium was formed in 1990, and started work in 1991 on the three-year task of producing a hundred-million word corpus of modern British English for use in commercial and academic research. The first edition of the corpus was published in 1994.

The project was funded by the commercial partners, the Science and Engineering Council (now EPSRC) and the DTI under the Joint Framework for Information Technology (JFIT) programme. Additional support was provided by the British Library and the British Academy.

The BNC Project is managed by the BNC Consortium and all major decisions regarding the BNC are still made by them. The Oxford University Computing Services (OUCS) acts as agent for the BNC Consortium in distributing the BNC under the terms of the User Licence.

Maintained by: BNC Webmaster *(bnc-queries@rt.oucs.ox.ac.uk)* .