British National Corpus 1994 to 2026

Martin Wynne, Megan Bushnell

University of Oxford
Faculty of Linguistics, Philology and Phonetics, Oxford, UK
{martin.wynne, megan.bushnell}@ling-phil.ox.ac.uk

Abstract

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. It is one of the first generation of monolingual, synchronic, general, representative corpora of its size, and led the way for other national corpora. It was created by a consortium of academic partners and publishers, with funding from the Department of Trade and Industry in the UK. This posters reflects on a number of lessons learning in more than thirty years, in terms of corpus representativeness, modes of access to the corpus, licensing, and managing the transition from a contemporary synchronic corpus to a historical corpus.

Keywords: Corpus linguistics, linguistic corpus, corpus construction, licensing

1. Extended abstract

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. It is one of the first generation of monolingual, synchronic, general, representative corpora of its size, and led the way for other national corpora. It was created by a consortium of academic partners and publishers, with funding from the Department of Trade and Industry in the UK[1].

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specialist periodicals and journals for all ages and interests, academic books and popular fiction, published and unpublished letters and memoranda, school and university essays, among many other kinds of text.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%) 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.

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)[2] 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. The corpus was an important landmark in the adoption of the TEI guidelines for a linguistic corpus. The creation of the corpus and its structural markup and annotation are fully documented in the User Reference Guide[3].

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 for copyright reasons prior to the release of the second edition BNC World (2001) and the third edition BNC XML Edition (2007). The part-of-speech tagging was revised and improved in the BNC Tag Enhancement project 1995-1996 at Lancaster University. The two-million-word BNC Sampler was manually annotated, then used to train the CLAWS part-of-speech tagger before automatically re-tagging the remaining c.98 million words of the corpus. The morphosyntactic annotations assigned at this time remain in all officially released versions of the corpus.

The BNC was originally made available on CD-ROM bundled with the Sara software for analysis and exploration of the text and tagging, with users paying a fee for media and administrative costs. Later, this was done on DVD with the Xaira (XML-aware) software, and in 2014 the BNC became available for download for free from the Oxford Text Archive, the CLARIN-UK repository. The Oxford Text Archive has been collecting corpora and other electronic texts and datasets since 1976, and celebrates its fiftieth anniversary in 2026.

A major project in the 2000s located a large proportion of the audio files on which the spoken corpus is based, aligned them with the text, and made them available from a streaming server[4]. A version of BNCWeb[5] makes use of this facility to offer the audio for concordance lines.

[1]https://www.natcorp.ox.ac.uk/corpus/creating.xml
[2]https://ucrel.lancs.ac.uk/claws/
[3]https://www.natcorp.ox.ac.uk/docs/URG/
[4]https://www.phon.ox.ac.uk/AudioBNC

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 76–77. Palma, Mallorca. May 11, 2026.

The licence for the BNC, agreed with the copyright owners of the materials included, only allows distribution on CD-ROM by Oxford University Computing Services on behalf of the BNC Consortium. This has been reinterpreted to allow online download, but only from the University of Oxford. However, the licence has been interpreted in such a way as to allow online corpus platforms to host the corpus and allow analysis and exploration, but not download of whole texts or the whole corpus. These platforms, including BNCWeb, English-Corpora.org and Sketch Engine, have been intensively used for many years. The BNC Licence is unusual in that it expressively allows and encourages commercial use of the corpus.

Researchers at Lancaster University created a comparable corpus BNC2014[6], a synchronic corpus of present-day English from a period 20 years after the original BNC. To facilitate identification and comparison of comparable corpora, the original BNC has been rebranded as BNC1994. One important lesson learned from the long history of the BNC, is that it is necessary to transition from branding a corpus as a snapshot of present-day language to a historical corpus. The corpus now also represents an important source of human language produced by native speakers from just before the internet age and the arrival of computer-mediated modes, and also from a time before the pollution of language data with computer-generated language.

2. Bibliographical References

BNC Consortium, British National Corpus 1994, Oxford Text Archive, http://hdl.handle.net/20.500.14106/2554

Coleman, J., Baghai-Ravary, L., Pybus, J., and Grau, S. (2012). Audio BNC: the audio edition of the Spoken British National Corpus. Phonetics Laboratory, University of Oxford. http://www.phon.ox.ac.uk/AudioBNC

Garside, R. (1996). The robust tagging of unrestricted text: the BNC experience. In J. Thomas and M. Short (Eds), Using corpora for language research: Studies in the Honour of Geoffrey Leech Longman, London, pp 167-180.

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)*. Kyoto, Japan, pp. 622-628.

Burnage, G. and Dunlop, D. (1993) Encoding the British National Corpus. In J. Aarts, P. de Haan and N. Oostdijk (Eds), English language corpora: design, analysis and exploitation. Amsterdam: Rodopi, pp. 79-95.

3. Language Resource References

BNC Consortium, British National Corpus 1994, Oxford Text Archive, http://hdl.handle.net/20.500.14106/2554

BNC Consortium, British National Corpus 1994 Sampler, Oxford Text Archive, http://hdl.handle.net/20.500.14106/2552

[5]http://bncweb.lancs.ac.uk/
[6]http://corpora.lancs.ac.uk/bnc2014/