BNC Baby v. 2 – three corpora on one CD

Ylva Berglund
Oxford University Computing Services (UK)

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 cross-section of late 20th century British English. The first release of the corpus was announced in February 1995. The corpus has since been revised and released in a second version (BNC World 2001). A new, third revised version in XML format is expected in 2006.

Although its size is one of the strengths of the BNC, it soon transpired that the 100-million-word collection was too big for certain uses. Using only a subset of the full corpus was always an option but many users felt that by creating a subset themselves they were missing out on the advantages of using a generally available collection that was easily obtainable by others. To cater for these users, and for the benefit of anyone who wants a smaller corpus with the features of the BNC, two sub-corpora of the full corpus have been created and released separately. These two sub-corpora, BNC Sampler and BNC Baby, are now available, together with the classic Brown corpus, on one CD: BNC Baby v. 2. All three corpora are in XML-format, ready to be used with the Xaira search program (included on the CD) or any other tool that handles texts in XML or plain text format.

BNC Sampler

The BNC Sampler corpus contains two million words in all, with equal proportions of written and spoken material. The texts were selected to mirror the composition of the full BNC corpus where possible. The word-class tagging of the BNC Sampler was done by the UCREL team at Lancaster, using a more detailed tagset than for the original BNC. The tagging was manually post-edited and corrected "and may be assumed to be almost error-free" (Leech and Smith 2000).
The BNC Sampler is particularly useful for those who want a corpus of a smaller size but with a mixture of texts of different kinds. It is also convenient

for comparisons between general written and general spoken language and a valuable resource for studies where the accuracy and consistency of the word-class annotation is vital.

The Sampler was first released in 1999 but has been out of print for several years. This new version contains the same texts and tagging as the original but has been converted to XML to be usable with the Xaira program.

BNC Baby

The BNC Baby consists of four one-million-word sub-corpora. The texts were selected from the BNC World corpus as examples of four different genres: fiction, newspapers, academic prose and spoken conversation (texts already included in the BNC Sampler were excluded). Known errors and typos in the texts were corrected. The BNC Baby sub-corpora have been tagged with word-class annotation and lemma information and converted to XML.

The BNC Baby has been used for studies where a smaller corpus of a particular kind of text is needed and also where variation between genres is in focus. It has been found useful for teaching, not least in combination with material discussing genre differences or corpus-based material, such as the *Longman Grammars* (Biber *et al.* 1999, 2002).

Brown Corpus

The third corpus on the BNC Baby v. 2 CD is the well-known one-million-word Brown corpus of American English. The corpus itself hardly needs any further introduction (see Francis and Kučera 1979). Just like the BNC sub-corpora, this classic corpus is made available with word-class tagging in XML format, ready to be used with the Xaira program.

Xaira

All the corpora on the BNC Baby v. 2 CD can be used with the Xaira corpus tool. The program is based on the SARA text searching software delivered with the BNC, but with many new features developed in response to feedback and criticism from SARA users. Xaira can be used on any corpus of well-formed XML documents.

Distribution

BNC Baby v. 2 CD is available from the Oxford University Computing Services and can be ordered via the BNC webpage http://www.natcorp.ox.ac.uk/. To encourage use in a classroom or multi-user environment, orders of 10 copies or more are offered a substantial discount.

References and further reading

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman grammar of spoken and written English.* Harlow: Pearson Education Limited.

Biber, Douglas, Geoffrey Leech, and Susan Conrad. 2002. *Longman student grammar of spoken and written English.* Harlow: Pearson Education Limited.

Burnard, Lou. 2000. *The British National Corpus users reference guide.* Available at http://www.natcorp.ox.ac.uk/docs/urg.html.

British National Corpus (BNC) website: http://www.natcorp.ox.ac.uk/.

Francis, W. Nelson and Henry Kučera. 1979. *Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with digital computers*. Available at the ICAME website: http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM.

Leech, Geoffrey. 1997. *A brief users' guide to the grammatical tagging of the British National Corpus.* Available at http://www.natcorp.ox.ac.uk/docs/gramtag.html.

Leech, Geoffrey and Nicholas Smith. 2000. *Manual to accompany the British National Corpus (version 2) with improved word-class tagging.* Available at: http://www.natcorp.ox.ac.uk/docs/bnc2postag_manual.htm.

University Centre for Computer Corpus Research on Language (UCREL) website, University of Lancaster http://www.comp.lancs.ac.uk/ucrel/.

Xaira website: http://www.xaira.org/.