BNC Design Model Past its Sell-By

Adam Kilgarriff, Sue Atkins, Michael Rundell
Lexicography MasterClass Ltd
{adam,sue,michael}@lexmasterclass.com

The BNC (British National Corpus) was an ambitious and innovative project. It has been intensively and extensively used in linguistics, language-teaching, language technology and dictionary-making. For researchers embarking on corpus research into other languages, the BNC, and in particular its design specification (Atkins *et al*. 1992) has often been taken as a model to work from.

This seems good practice: the BNC was a well-thought-through, highly successful project, so others should use it as a model.

However the BNC was designed seventeen years ago. It is pre-Web.

The Web changes a premise on which the BNC model was based. When the BNC was planned, its 100 million words made it far, far larger than existing corpora. Its prime movers were dictionary publishers, and they knew they wanted as large a corpus as they could possibly get. 100 million was, in 1990, that dream.

Seventeen years on, 100m words is commonplace. Google gives us everyday access to *eighty thousand* times as much. 100m word corpora can be built to order in a few weeks (see http://corpus.leeds.ac.uk/internet.html) Thus the vision behind the BNC translates, in 2007, to a completely different corpus.

The BNC vision had other aspects besides the (at the time) mind-boggling size:

1. a balance of text types
2. a substantial share (10%) of spoken language
3. no debilitating copyright constraints
4. a reference corpus

How does each of these look now?

1. The balance of text types in the BNC was selected by a committee comprising dictionary publishers and academics. While their inventory of text types, and proportions, were thoroughly reasonable, they were not and could not have been objective, and should not be given undue weight. They were constrained by how costly it would be to gather each type. The costs are now radically changed, falling to virtually zero for online newspapers and blogs (which did not exist then).

2. For spoken language, the availability of transcribed material online (see eg Hoffman 2007) and automatic transcription (see eg http://podzinger.com) change the landscape, for English with other major languages probably to follow.

3. The most painful part of creating the BNC was obtaining permission from authors and publishers so that the corpus could be distributed. Even so, it is still constrained in what can be done with it (in contrast to, say, WordNet). In 2007, it would still be desirable, for any language, to have a corpus which could be copied without constraints. The copyright status of the web is a legal quagmire, with Google and Yahoo copying billions of documents everywhere, every day, but with legal cases such as Napster casting a shadow.

4. A 'reference corpus' is a corpus that people can use as a reference point for the language. For that it should be balanced (see 1 above) and freely available (see 3 above). Size is not always important: the Brown corpus has been doing sterling service for some types of study for forty years. What a reference corpus ideally contains (and how big it needs to be) depends on the use to which it will be put. Getting the 'right' reference corpus from the Web will often be better than using whatever happens to be available. (This has been our experience with terminology-finding in WebBootCaT (Baroni *et al*. 2006).)

The BNC was a large project with a budget of over a million pounds. In 1990, the expenditure was entirely justified. The BNC met very many otherwise-unmet needs. In 2007, many of these are already met, mostly by developments related to the Web, for many languages. The BNC model was successful in its time, but that does not imply it is appropriate today.

References

Atkins, S, J. Clear and N. Ostler (1992). Corpus Design Criteria. Literary & Linguistic Computing 7:1:1-16

Hoffman, S. (2007). From Web-Page to Mega-Corpus: The CNN Transcripts." In: Hundt, Nesselhauf and Biewer (eds.) *Corpus Linguistics and the Web.* Amsterdam: Rodopi. 69-85.

Baroni, M, A Kilgarriff, J. Pomikálek and P. Rychlý (2006). WebBootCaT: instant domain-specific corpora to support human translators. Proc EAMT Workshop on tools for human translators. Oslo, Norway.