The British National Corpus Revisited: Developing parameters for Written BNC2014

Abi Hawtin (Lancaster University, UK)

1. The British National Corpus 2014 project

The ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University and Cambridge University Press are working together to create a new, publicly accessible corpus of contemporary British English. The corpus will be a successor to the British National Corpus, created in the early 1990s (BNC1994[1]). The British National Corpus 2014 (BNC2014) will be of the same order of magnitude as BNC1994 (100 million words). It is currently projected that the corpus will reach completion in mid-2018. Creation of the spoken sub-section of BNC2014 is in its final stages; this paper focuses on the justification and development of the written sub-section of BNC2014.

2. Justification for BNC2014

There are now many web-crawled corpora of English, widely available to researchers, which contain far more data than will be included in Written BNC2014. For example, the English TenTen corpus (enTenTen) is a web-crawled corpus of Written English which currently contains approximately 19 billion words (Jakubíček et al., 2013). Web-crawling processes mean that this huge amount of data can be collected very quickly; Jakubíček et al. (2013: 125) note that "For a language where there is plenty of material available, we can gather, clean and de-duplicate a billion words a day." So, with extremely large and quickly-created web-crawled corpora now becoming commonplace in corpus linguistics, the question might well be asked why a new, 100 million word corpus of Written British English is needed.

2.1 The enduring popularity of BNC1994

One answer to that question is the demonstrable value even two decades on of BNC2014's earlier counterpart, BNC1994. Despite being created in the 1990s and containing data from as far back as the 1960s, BNC1994 is still extremely widely used in linguistic research. This is perhaps surprising, because BNC1994 no longer represents contemporary British English, and is certainly not anywhere near the largest available corpus of British English.

A simple search for the term "British National Corpus BNC" in Lancaster University's online library catalogue yields 935 results (although this figure does include some repeats). Just under half of these results were works published from 2010 onwards. This shows that BNC1994 continues to be a very productive data source for research right up to the present day.

[1] The name 'BNC1994' is not widely used, but the decision has been made to refer to the corpus in this way in order to make the link between BNC1994 and BNC2014 clear.

So, if BNC1994 is not the largest available British English corpus, and is also not the most contemporary, there must be a different reason why BNC1994 continues to be so productive for linguistic research. To appreciate this, it is informative to revisit some of the stated goals of the creators of BNC1994:

- To create a corpus an order of magnitude larger than any currently freely available corpus.
- To create a synchronic corpus.
- To include a range of samples from the full range of both spoken and written British English.
- To create the corpus using a non-opportunistic design.
  (Burnard, 2002: 53).

I argue that the reason that no large, web-crawled, or more contemporary corpora have enjoyed the level of uptake of BNC1994 is because none of them meet all of these goals in the way that BNC1994 does. Many corpora, such as BE06 (Baker, 2009), are more contemporary than BNC1994 but are much smaller, whilst many other corpora, such as enTenTen (Jakubíček et al., 2013), are more contemporary and much larger, but do not meet the goal of being synchronic, cannot guarantee the language contained in them is British English, and also do not contain samples from the full range of British English.

2.2 The benefits of a 'hand-made' corpus

We see then, certain benefits which BNC1994 has over large web-crawled corpora; let us therefore consider in more detail specific examples of how the 'hand-made' nature of Written BNC2014 will set it apart from other corpora. By 'hand-made', I mean a corpus where texts are selected by manual procedures, perhaps assisted by automatic measures, but with the choice of texts ultimately made by a human and not by unsupervised software as is normally the case for general-purpose web-crawled corpora.

One of the benefits of a 'hand-made' corpus is that this method allows the corpus creators to have much greater control over what texts are included, thus guaranteeing with a rather higher level of certainty than in web-crawled corpora that the selected texts were written by speakers of British English. Jakubíček et al. (2013: 126) trained a classifier to distinguish between British and American English when creating enTenTen; such an approach, however, is less rigorous than a human manually ascertaining this information. Of course, to accomplish this effectively for a very large number of texts (approximately 20,000) appropriate procedures and strategies are required, the development of which is a present focus of the BNC2014 project effort.

A further benefit of a 'hand-made' corpus is that we will be able to include within BNC2014 data types which a web-crawl cannot – chiefly, published books. A significant amount of time will be dedicated to negotiating with publishers to include published books within Written BNC2014, which will set the corpus apart from many other contemporary corpora. It is planned that Written BNC2014 will contain 41 million words of published books.

Of course this time spent manually collecting certain texts will result in the need to partly automate large areas of data collection in order to assist the manual effort. Our approach to periodicals in particular is to perform mass-downloads from the web using crawling techniques, but then, critically, for project personnel to work manually with this data, and thereby to select appropriate texts for inclusion.

Given the time investment which will be required to seek publisher's permission to sample extracts of published books, we have determined that for all other kinds of writing, we will not attempt to seek copyright holders' permission, but will instead take advantage of certain relevant exceptions in UK copyright law (UK Copyright Service, 2015). Written BNC2014 will be a non-commercial project, thus, under the 'Non-commercial research' exception, we will not breach any intellectual copyright law in "copy[ing] limited extracts of works" (UK Government, 2014). Such use must be within 'fair dealing' and must not lead to any financial impact on the copyright holder (UK Government, 2014). On the latter point, we consider it is highly unlikely that there would be any financial impact on any of the copyright holders of works included in Written BNC2014, because the eventual texts will be so heavily transformed, with XML markup and word-level annotation, that it is doubtful that anyone would try to read the text in Written BNC2014 rather than the original. 'Fair dealing', meanwhile, is "a legal term used to establish whether a use of copyright material is lawful or whether it infringes copyright" (UK Government, 2014). Fair dealing is determined on a case-by-case basis, and factors which have been deemed relevant by courts in determining fair dealing include whether the use of the work affects the market for the original, and whether the amount of the work used was reasonable, appropriate, and necessary (UK Government, 2014). The use of works in Written BNC2014 is highly unlikely to affect the market for the original, and, as only small samples (of around 5,000 words[2]) will be taken from texts, this use is likely to be considered to fall within the limits of fair dealing. Approaching data collection under these exceptions to UK copyright law will give the benefit of a corpus which contains a broad mix of carefully selected texts which we could not have achieved without using these exceptions.

3. Conclusion

In conclusion, it is very likely that the British National Corpus 2014 will prove as valuable to the research community as its predecessor, because it will have features simply not present in large web-crawled corpora. This is not to say that web-crawled corpora are not themselves valuable resources for particular purposes, most notably purposes for which the combination of size and speed of collection is especially desirable. A corpus of the size that will take the BNC2014 project team years to create could be created by a web-crawler in a matter of hours. However, smaller 'hand-made' corpora such as BNC1994 and BNC2014 have in their own sphere equally many advantages; as this paper has illustrated, our design decisions and procedures for corpus construction have been defined with these particular strengths of BNC-style corpora in mind.

[2] 2 This is in stark contrast to BNC1994 where texts were often tens of thousands of words in length, and which contains only 4049 texts.

References

Baker, P. (2009). The BE06 corpus of British English and recent language change. *International Journal of Corpus Linguistics*, 14(3), 312-337.

Burnard, L. (2002). A retrospective look at the British National Corpus. In: B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp.51-72). Amsterdam – New York: Rodopi.

Davies, M. (2013b). Google Scholar and COCA-Academic: Two very different approaches to examining academic English. *Journal of English for Academic Purposes*, 12(3), 155-165.

UK Government (2014). *Intellectual property – guidance. Exceptions to copyright*. Retrieved from: https://www.gov.uk/exceptions-to-copyright.

ICAME36 (2015). *Words, words, words – Corpora and lexis, ICAME36 Abstract book*. University of Trier.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen Corpus Family. In A. Hardie & R. Love (Eds.), *Corpus Linguistics 2013 Abstract Book* (pp. 125-127). Lancaster: UCREL.

UK Copyright Service (2015). *UK Copyright Law*. Retrieved from: http://www.copyrightservice.co.uk/ukcs/docs/edupack.pdf.