Institution: University of Oxford

Unit of Assessment: 28 Modern Languages and Linguistics

Title of case study:

Public dissemination of the British National Corpus

1. Summary of the impact (indicative maximum 100 words)

The 100-million-word British National Corpus of UK English texts and speech is used regularly and extensively as a reference resource on the contemporary English language. Its users include dictionary makers, school teachers in many countries, teachers of English as a second language, the OCR school examinations board, and many individual writers on the internet as a reference source about questions of contemporary English usage. Its use has led to improved English dictionaries that more accurately reflect actual usage: for Longman, Chambers and OUP dictionaries, use of the BNC provides a unique selling point over their competitors and enhanced educational value to readers. For students and English language teachers world-wide, the BNC provides more realistic examples of the usage of words and phrases in context, and in different registers, free of charge via various online search portals, and thus improved education in English.

2. Underpinning research (indicative maximum 500 words)

Variation is rampant in human language across speakers, gender differences, social classes and dialects. Language variation and use and the phonetic investigation of variation in speech are fundamental parts of linguistic research. Spoken language differs in important ways from the formal written language that underlies many dictionaries and language teaching material. The corpus was collected by the BNC Consortium, an industrial/academic partnership initiated by Oxford University Press, joined by two other dictionary publishers, Longman and Chambers, together with corpus linguistics researchers from Oxford University, Lancaster University, and the British Library's Research and Innovation Centre.

Lou Burnard (Applications Programmer Manager until 1995; Manager, Humanities Computing Unit 1995-2001; Assistant Director, Computing Services, 2001-2010; retired 9/2010) was responsible for Oxford University's participation in the consortium project, especially the design of its encoding scheme, and subsequent curation and worldwide distribution of the corpus. Professor John Coleman (Director of the Phonetics Laboratory at the University of Oxford since 1993) was responsible for a collaboration with the British Library to digitize the spoken recordings, and for two successive projects that have aligned the transcriptions to the audio, anonymized the relevant portions, and published the speech in its entirety.

The initial iteration of the BNC was created by an academic/industrial consortium with joint DTI/SERC funding until 1994 with a second edition published in 2001, a major revision of the corpus produced a third (XML), more readily-accessible edition in March 2007. [Reference 1]. Oxford's roles in the consortium were linguistic structural mark-up of the texts (collected and supplied by the publishers) in standard formats (SGML; later, XML following the conventions of the Text Encoding Initiative), validation of its consistency, documentation, and publication. Part-of-speech tags were provided by Lancaster University. Oxford University Press, Longman and Chambers obtained and digitized the text samples and managed their IP/copyright permissions, and Longman collected or arranged for collection and transcription of the spoken audio recordings, on cassette tapes. The British Library curated those tapes which, until 2011, could only be audited by visiting their Sound Archive in person.

In 2009-2013, the British Library digitized the original audio recordings of the 10-million word spoken part; Oxford University Phonetics Laboratory played the crucial role of automatically aligning them with their transcriptions, and published the anonymized audio and time-aligned transcriptions on-line [Reference 2]. To enable modern web browsers to access any desired audio clip from any of the recordings, the Laboratory implemented server-side audio fragment streaming. Because it is a well-documented on-line publication, users easily find it by generic search engine enquiries such as “spoken BNC”. It is an excellent source of data for natural language processing and speech technology and commercial communications research. Oxford researchers have also published on-line and print reference and handbook documentation [References 3, 4].

3. References to the research (indicative maximum of six references)

Key outputs from the research described in the previous section

[1] The BNC Consortium (2007) *BNC XML Edition*. Can be obtained in its entirety on DVDs from http://www.natcorp.ox.ac.uk/getting/index.xml?ID=order, or inspected from http://www.natcorp.ox.ac.uk or several other portals.

[2] John Coleman, Ladan Baghai-Ravary, John Pybus, and Sergio Grau (2012) *Audio BNC: the audio edition of the Spoken British National Corpus. Phonetics Laboratory, University of Oxford*. http://www.phon.ox.ac.uk/AudioBNC.

[3] Lou Burnard, ed. (2007) *Reference Guide for the British National Corpus (XML Edition)* http://www.natcorp.ox.ac.uk/docs/URG/.

[4] Lou Burnard and Guy Aston (1998) *The BNC Handbook: exploring the British National Corpus with SARA.* Edinburgh: Edinburgh University Press. ISBN: 9780748610556 http://www.natcorp.ox.ac.uk/archive/sara/index.xml http://www.euppublishing.com/book/9780748610556

The quality of [1]–[3] is evidenced by the fact that their creation and development was funded by the Department of Trade and Industry, SERC (1990-1994), JISC (as part of the Digging into Data competition – a partnership of UK, US and Canadian research funding agencies; £100,000 from JISC 1/1/10-30/6/11), and ESRC (£543,700,1/11/10-31/10/13) and by the quality and prestige of the partners in the consortium i.e. Longman, Chambers and the British Library. [4] was published by a University Press.

4. Details of the impact (indicative maximum 750 words)

The impact claimed in this study arises from the production and public availability of the corpus's texts and spoken audio in a variety of media – CD-ROM, DVD, and on-line – together with associated documentation, publications and access software. The consortium also published SARA, a concordance tool for search and analysis of the BNC.

An estimated 2000 full licences (both personal and institutional licences) and 1800 copies of BNC Baby (a 4 million word sample of the full BNC) have been sold world wide since 2008. In addition, our Japanese distributor has distributed over 350 licences in the relevant period. **[§5i]**

The impacts have been:

1) A contribution to **United Kingdom's cultural heritage**. A century ago Oxford University researched and published the authoritative Oxford English Dictionary, based upon written texts. The 100-million-word British National Corpus (10 million spoken words) is a documentary record of written and spoken English representing a wide cross-section of British English, at the end of the 20th century.

2) For **dictionary makers**, the BNC improves and adds value to their products, providing a unique selling point over their competitors, and enhancing educational value to purchasers of their dictionaries. The corpus was used by the dictionary publishers making up the original consortium: OUP, Longman and Chambers, in creating the Longman Dictionary of Contemporary English, the Chambers 21st Century Dictionary, and the Oxford Advanced Learner's Dictionary.

- Longman Dictionaries' website states: “What is so unique about the Spoken Corpus is that it shows us how we really use English, not how we are supposed to use English or how we use it when we are writing. It reveals how very different the spoken word is from the written word. At last students will be able to study English in an exciting new range of ELT materials that represent English as it really is. For the first time, real spoken language has influenced the creation of a learner dictionary.”
- Chambers 21st Century Dictionary claims “All Chambers dictionaries are supported by the British National Corpus, a 100-million-word database of written and spoken English that provides real evidence of how English is used.”
- The Oxford Advanced Learners' Dictionaries website says “We use the BNC to confirm our intuitions and also to tell us things we didn't already know, or may not have thought about. We can find out exactly what a word means, rather than what we think it means. We can see how it

behaves grammatically and which words it collocates with. We use all this information when writing our learners' dictionaries."

Since 2011, the Government of India (Central Institute of Indian Languages) has been collaborating with Longman Pearson to create bilingual dictionaries in English and eleven Indian languages. The English part has been developed using words and phrases culled from the BNC. According to the Project manager of the National Translation Mission "The lexicon is an important translation tool - kind of a spring board to push the mother tongues, many of which are threatened with very few speakers". **[§5ii]**

David Crystal's chapter 'Essential Grammar' for English language learners in The Longman Essential Activator draws on the BNC for examples of real-life English usage to help intermediate learners to better understand how to use grammar in written and spoken English. **[§5iii]**

3) **English language teachers** world-wide, and the agencies that examine them, have better examples of contemporary spoken English to use in their studies, improving the standards of teaching and learning of English as a Foreign Language.

In comments sent to researchers via the BNC registration page, Juan, an English teacher in Spain says he uses the recordings in his class so that his students become aware of the different accents in the UK. Another teacher uses it "for my students to practise repeating after the tapes to improve their pronunciation." Tdol, an English teacher in Japan, refers language users to the BNC (20/3/2008). A blog post by Monica Vlad, a language teacher from Romania recommends language learners consult the BNC to check on questions of usage "in a few easy steps without having to google endlessly and end up with some website or forum results — you can't really know how reliable they are."**[iv]** Other such comments from BNC user registrations demonstrate the reasons they are using it and the impact it has on improving teaching and learning of English as a foreign language:

- "Teaching English to Swedish students, exposing them to a range of accents."
- "As a teacher of English, for personal reasons, so as to check doubts about collocation, use... but also for professional reasons."
- "I am in English teacher in Spain and would like to use the recordings in my class so that my students become aware of the different accents in the UK."
- "I could use [the recordings] as authentic listening exercises especially with regard to different accents."
- "looking for recording by natives for my students"

4) The BNC provides **English language learners** with access to copious audio and text examples from an authoritative source which they can use to become familiar with variation in English and to improve their own British English pronunciation and usage, supporting independent study. The BNC site registration records provide evidence that the spoken audio is used by learners for purposes such as "to know how to pronounce words and practice how to listen them within the context", "to refresh my English", "to learn more about British English", "to know how to pronounce words and practice how to listen them within the context", "just listen to refresh my English; it has been too long since I could talk to native speakers and listen to spoken English at large" and "I want to improve my English by listening to these audio files".

5) The OCR **school examinations board** are using extracts of BNC conversation transcriptions originally published on the Phonetics Laboratory website in order to improve the quality of public examinations, and since March 2013 have been selling copies of the exam papers.**[§5v]** Analysis of natural English conversation is a key part of the English Language GCSE and A-level curricula.

6) People all over the world with **access to the internet** use the BNC as an authoritative reference on English, without having to pay for it or buy dictionaries or other reference works. This has widened access to accurate information about present-day UK English. Many individual writers use BNC as a reference source relevant to questions of contemporary English usage. It is a superior resource in comparison with other online sources: Jonathan de Boyne Pollard claims BNC can be a more accurate source of word frequencies than Google hit counts. **[§5vi]**

To demonstrate an estimate of the internet reach of the BNC, the search term British National Corpus yields 2.18 million hits on Google UK (cf. the 1.3m hits for their closest UK competitor, the Edinburgh

map task corpus. The Google search engine lists over 80,000 blog posts and over 17,000 discussions citing British National Corpus. However, large numbers of users look up words or phrases via various online search services, especially the public, simple search engine provided by the British Library on the main BNC site. According to Google there are currently about 18,300 incoming links to that site and about 1,330 direct links to the search engine. **[§5vii]**

Other online language tools use the BNC as a basis Sharp Laboratories of Europe has created an on-line writing aid, 'Just the Word' based upon BNC usage data. One user writes "I use it every day to check the combinability of a word, to see the context and get the feel of a word…It is such a great help!!! …You are doing a tremendous job making a difference to those whose native language is not English." **[§5viii]**

7) BNC is used as a source of data for **natural language processing and speech technology** industry e.g. Infochimps is a start-up company founded in 2010 that is making a market for people to buy and sell large datasets. This includes a dataset of word frequencies from the BNC available for download under a Creative Commons licence. **[§5ix]**

8) People have used the BNC for more **artistic applications**: it was used as the source for "Found Poetry" by the Three Poets Tumblr blog. The poets pick a word and search for it in the BNC and create a poem from the results. **[§5x]** WordCount is an "online artistic experiment in the way we use language" based on the BNC. It presents the 86,800 most frequently used English words, shown visually scaled in order of commonness. Barbara Wallraff cited WordCount and BNC in an article about vocabulary use in 2008. **[§5xi]** It was also cited on the 'Shape+Colour' blog by Jeremy Elder (2008).

5. Sources to corroborate the impact (indicative maximum of 10 references)

[i] Sales figures for BNC from Oxford University Stores records
http://www.oxforduniversitystores.co.uk/browse/category.asp?compid=1&modid=1&catid=1049

[ii] Quote taken from 'Bilingual dictionaries to promote India's mother tongues', 12/03/12, TwoCircles.net
http://twocircles.net/2012mar12/bilingual_dictionaries_promote_indias_mother_tongues.html

[iii] Crystal (2006) Essential Grammar. In *The Longman Essential Activator* 2nd edition, pp911-936.

[iv] Monica Vlad, 'How to use corpora for English language teaching and learning', APLaNet blog, 23/05/11, http://aplanet-project.org/profiles/blogs/how-to-use-corpora-for-english

[v] OCR Copyright acknowledgement booklet for the January 2013 exam series , pp18-19
http://www.ocr.org.uk/Images/129661-copyright-acknowledgement-booklet-january-2013.pdf

[vi] Jonathan de Boyne Pollard, 'Google result counts are a meaningless metric', 2008.
http://homepage.ntlworld.com./jonathan.deboynepollard/FGA/google-result-counts-are-a-meaningless-metric.html

[vii] Source of figures is the google.co.uk search engine results counter and the BNC webmaster. The Google figures change continuously but give some idea of the order of magnitude of results.

[viii] *Just the Word* user comment - Elfina, 'Thank you for your amazing work!', 02/06/2010
https://groups.google.com/forum/#!topic/just-the-word-users/PjuANtoLf1A

[ix] Geoffrey Leech, Paul Rayson, Andrew Wilson, Word Frequencies in Written & Spoken English from British National Corpus (100M-word), Infochimps Datasets blog, Post Created about 3 years ago http://www.infochimps.com/datasets/word-frequencies-in-written-spoken-english-from-british-national

[x] Three Poets, 'Found Text Poem 1', 19/04/13, Three Poets Tumblr blog
http://threepoets.tumblr.com/post/48375193990/found-text-poem-1

[xi] Barbara Wallraff, 'In a Word' *The Atlantic Monthly,* 302.4, 01/11/2008, p.142.
http://www.theatlantic.com/magazine/archive/2008/11/in-a-word/307067/