Introduction

Compiling and analysing the Spoken British National Corpus 2014

Tony McEnery, Robbie Love and Vaclav Brezina
Lancaster University

For over twenty years, the British National Corpus has been one of the most widely known and used corpora. It is almost impossible to attend an international corpus linguistics conference such as Corpus Linguistics, ICAME (International Computer Archive of Modern and Medieval English), AACL (American Association for Corpus Linguistics) or APCLC (Asia Pacific Corpus Linguistics Conference) without encountering several papers which in some way employ the BNC. Focusing on the 10-million-word spoken component of the BNC, Love et al. (this issue) show that no other orthographically transcribed spoken corpus compiled since the release of the BNC has matched the Spoken BNC in either its size or availability. Unsurprisingly, the corpus linguistics community has, for some time, used the Spoken BNC as a proxy for "present-day" spoken British English. That the "go-to" dataset is over twenty years old, as Love et al. (this issue) argue, is a problem for current and future research that needs to be addressed with increasing urgency.

The collaboration between Cambridge University Press (CUP) and the ESRC Centre for Corpus Approaches to Social Science (CASS)[1] at Lancaster University to build the Spoken BNC2014 came about after some years of both centres working individually on the idea of addressing this situation by compiling a new corpus of spoken British English which could, in some way, match up to the Spoken BNC.[2] Claire Dembry at CUP had collected two million words of new spoken data for the Cambridge English Corpus between 2012 and 2014, trialling the public participation method which was retained, along with the data itself, in the Spoken BNC2014

1. The research presented in this paper was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1.

2. CASS is also working on the development of the Written BNC2014 (with Abi Hawtin investigating written corpus construction), which will be publicly released at a later stage.

International Journal of Corpus Linguistics 22:3 (2017), **311–318**. DOI 10.1075/ijcl.22.3.01mce
ISSN 1384-6655 / E-ISSN 1569-9811

(see Love et al. this issue). Meanwhile, Tony McEnery and Andrew Hardie at Lancaster had been planning to compile a new BNC and, by 2013, had recruited Robbie Love to start investigating methodological issues in compiling spoken corpora, and Vaclav Brezina, to bring insights to the project based on his use of the Spoken BNC1994 to explore sociolinguistic research questions. Early in 2014, both parties agreed, upon learning of each other's work, to pool resources and work together to build the 'Lancaster/Cambridge Corpus of Speech' (LCCS) which, within a few months and with the blessing of Martin Wynne at the University of Oxford, was renamed the Spoken British National Corpus 2014 (Spoken BNC2014).[3]

The Cambridge team was responsible for the gathering and transcription of recordings, while the Lancaster team would convert the resulting texts into an appropriate version of XML and annotate these files for hosting on Lancaster University's *CQPweb* server (Hardie 2012). Both teams collaborated on issues such as participant recruitment, speaker and recording metadata, design, ethics and transcription conventions.[4]

Once the project was under way and an anticipated release date of September 2017 was established, the team invited proposals from scholars who wished to gain exclusive early access to five million words of the data to conduct a research project of their choosing. This sub-corpus, known as the Spoken BNC2014 Sample (Spoken BNC2014S), is less than half the size of the full version of the corpus and contains transcripts from conversations recorded between 2012 and 2015 (its structure, which differs from the full corpus, is described in the Appendix). The Spoken BNC2014S was made available to the authors of eleven successful proposals. These were selected based on their innovative use of the data and the significance of the topic; the authors represent a cross-section of early-, mid- and late-career researchers from around the world. Four of the resulting research papers are featured in the current issue, while the remaining ones will be published in a forthcoming book with a focus on sociolinguistic variation (Brezina et al. forthcoming). The publication of both this special issue of *IJCL* and of the book are intended to celebrate the public launch of the Spoken BNC2014 and demonstrate some of its uses.

And so we gladly turn our attention to the current special issue of *IJCL*, in which we present a selection of work which demonstrates the usefulness of the new

---

3. In turn, the original Spoken BNC was retrospectively named the Spoken BNC1994 by the research team to distinguish it from its successor.

4. Love et al. (this issue) describes in greater detail how the Lancaster/Cambridge partnership designed and built the Spoken BNC2014, and the BNC2014 user guide (Love et al. 2017) includes information about the structure of the full 11.5-million-word corpus.

dataset and, in some cases, gives a snapshot of possible changes in spoken British English between the 1990s and the 2010s. The first paper more thoroughly introduces the Spoken BNC2014, while the remaining four delve into the data, with intensifiers, verb-forming suffixation, demonstrative clefts and downtoners in focus.

Love, Dembry, Hardie, Brezina and McEnery describe the method used to compile the Spoken BNC2014. The underlying theme of the paper is the maximisation of the efficiency of spoken corpus creation in view of practical constraints, with the focus on the principles of design as well as data and metadata collection, transcription and processing. As is not unusual in corpus construction, compromises had to be made throughout the compilation of this corpus; these are laid out transparently. Furthermore, the paper describes the innovative aspects of the Spoken BNC2014 project – notably including the use of PPSR (public participation in scientific research, Shirk et al. 2012), the introduction of new speaker metadata categorisation schemes, and consideration of the difficulty of speaker identification at the transcription stage. While the paper does not attempt to function as a Spoken BNC2014 "user guide" (cf. Love et al. 2017), it is a thorough account of the careful decisions that were made at each stage of development, and should be read by anyone who uses the corpus.

Fuchs investigates how age, gender, socio-economic status and dialect influence the use of intensifiers, and how this may have changed between the demographically-sampled (DS) component of the Spoken BNC1994 and the Spoken BNC2014S. All four sociodemographic factors are found to influence the use of intensifiers. Most notably, male speakers are found to use intensifiers less frequently than female speakers in most age groups and socio-economic groups; however, the use of intensifiers has risen across the board over time. Furthermore, gender differences in the intensifier use appear to have diminished to some extent – especially in the middle class. The paper provides a good demonstration of the capability to compare the Spoken BNC2014 with its predecessor for the purpose of investigating sociolinguistic variation and change in spoken British English between the 1990s and the 2010s.

Laws, Ryder and Jaworska are interested in the process of verb formation using derivational morphology – specifically via four principal verb-forming suffixes in English: *-ate*, *-en*, *-ify*, and *-ize*. By comparing the Spoken BNC1994DS with the Spoken BNC2014S, they examine the effects of speaker age and gender on lexical diversity, density and creativity, and how these may have changed over time. As predicted, younger age groups are found to exhibit a more restricted range of verbs in both time periods. Male speakers are found to use a wider repertoire of complex verb forms (types), but, contrary to previous research, token frequency counts are not found to associate with gender. Their contribution further demonstrates the

new sociolinguistic research possibilities afforded by comparison between the two Spoken British National Corpora.

Calude's study is concerned with demonstrative clefts (e.g. *that's what I wanted to talk about*), and aims to document sociolinguistic patterns of variation in the Spoken BNC2014S. Analysing nearly 6,000 demonstrative cleft constructions, logistic regression tests show that older adults (30–59 years) use them significantly more than children and young adults (1–29 years); in addition, speakers with schooling prefer them to those without and male speakers are more likely to use them than female speakers. Furthermore, speakers are reported to prefer *that*-demonstrative clefts as opposed to *this*-demonstrative clefts – especially so for speakers in the middle class as opposed to those in the highest socio-economic group. The paper provides not only a thorough discussion of the sociolinguistic implications of such findings but it also pays careful attention to methodological rigour, for example, in the treatment of outlier speakers and aggregate data. Overall, it shows how the Spoken BNC2014 may be used to study spoken language at the grammar and discourse level under a variationist lens.

Hessner and Gawlitzek, like Fuchs, are interested in intensifiers, but are exclusively interested in gender as opposed to other sociodemographic variables. Furthermore, they restrict the scope of their analysis to consider the Spoken BNC2014S only. They start by presenting frequency-based results which agree with the finding of Fuchs that female speakers use intensifiers significantly more often than male speakers. Beyond this, by considering amplifiers and downtoners separately, the study shows that this difference is caused only by changes in the use of amplifiers; downtoners do not differ significantly according to gender. While Fuchs' paper presents a bird's eye view of intensifier use across sociodemographic groups and time periods, this paper nicely shows how the Spoken BNC2014 might be used to explore language in context in more detail (e.g. by studying collocation).

As the brief summaries show, all the papers in the current special issue are clearly linked by their description and use of a brand-new dataset – the Spoken BNC2014. While these papers are necessarily few in number, it is our hope that they nonetheless demonstrate just some of the many applications for which the corpus will be used now that it has been released. We would, of course, like to thank Michaela Mahlberg for providing this opportunity to introduce the Spoken BNC2014. We are also very grateful to the reviewers of this special issue who provided peer review for each of the papers with skill and generosity. In turn, the authors of these papers, who have been in contact with us for the last two years and cooperated so competently, were able to respond to the comments provided by the reviewers in timely fashion and with a tight deadline. Furthermore, the authors' use of the Spoken BNC2014S trialled several new features of the *CQPweb* interface,

and allowed us to experiment with some new ideas for the categorisation of speaker metadata (e.g. age). We are very grateful to the authors for their comments and suggestions for improvements to these features and categorisations, which have been implemented for the full release of the Spoken BNC2014. We would also like to thank all others who submitted proposals for early access to the data, who showed enthusiasm about the project and will now enjoy access to the full version. Finally, we wish to thank Gavin Brookes and Lorenzo Mastropierro for their superb editorial assistance at all stages of the compilation of this special issue of *IJCL*.

References

Brezina, V., Love, R., & Aijmer, K. (Eds.) (forthcoming). *Corpus Approaches to Sociolinguistic Variation in Contemporary British English: An Exploration of the Spoken BNC2014*. New York: Routledge.

Hardie, A. (2012). CQPweb – Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics, 17*(3), 380–409. doi:10.1075/ijcl.17.3.04har

Love, R., Hawtin, A., & Hardie, A. (2017). *The British National Corpus 2014: User Manual and Reference Guide (version 1.0)*. Lancaster: ESRC Centre for Corpus Approaches to Social Science.

Shirk, J. L., Ballard, H. L., Wilderman, C. C., Phillips, T., Wiggins, A., Jordan, R., McCallie, E., Minarchek, M., Lewenstein, B. V., Krasny, M. E., & Bonney, R. (2012). Public participation in scientific research: A framework for deliberate design. *Ecology and Society*, 17(2), 29. doi:10.5751/ES-04705-170229

Appendix. Population of the main speaker demographic categories in the Spoken BNC2014 Sample (BNC2014S)

Gender

Gender	No. words
F	2,872,758
M	1,911,836
X	97
Total	**4,784,691**

Age

Age	No. words
0–10	1,281
11–18	191,987
19–29	1,961,779
30–39	834,379
40–49	463,022
50–59	375,368
60–69	625,013
70–79	254,263
80–89	45,066
90–99	3,812
Unknown	28,271
Total	**4,784,241**

Socio-economic status: NS-SEC

NS-SEC	No. words
1.1	81,728
1.2	106,0691
2	1,498,777
3	527,335
4	95,523
5	93,005
6	78,227
7	40,390
8	668,608
Uncat	614,721
Unknown	25,687
Total	**4,784,692**

Socio-economic status: Social Grade

SG	No. words
A	1,142,419
B	1,498,777
C1	622,858
C2	93,004
D	118,617
E	1,283,329
Unknown	25,687
Total	**4,784,691**

Region

Global	Country	Supra-region	Region
UK (4,419,193)	English (4,358,132)	North (1,158,231)	North East (320,464)
			Yorkshire & Humberside (478,268)
			North West (not Merseyside) (155,552)
			Merseyside (116,420)
		Midlands (375,259)	East Midlands (28,178)
			West Midlands (58,880)
			Eastern (378,065)
		South (2,470,535)	South West (33,104)
			South East (not London) (215,420)
			London (188,188)
	Scottish (10,440)	Scottish (10,440)	Scottish (10,440)
	Welsh (40,843)	Welsh (40,843)	Welsh (40,843)
	Northern Irish (0)	Northern Irish (0)	Northern Irish (0)
Non-UK (74,214)	Irish (12,462)	Irish (12,462)	Irish (12,462)
	Non-UK (61,752)	Non-UK (61,752)	Non-UK (61,752)
Unspecified (291,284)	Unspecified (301,062)	Unspecified (655,169)	Unspecified (2,686,655)

Authors' addresses

Tony McEnery
Centre for Corpus Approaches to Social Science (CASS)
Faculty of Arts and Social Sciences
Lancaster University
Lancaster, LA1 4YD
UK
a.mcenery@lancaster.ac.uk

Robbie Love
Centre for Corpus Approaches to Social Science (CASS)
Faculty of Arts and Social Sciences
Lancaster University
Lancaster, LA1 4YD
UK
r.m.love@lancaster.ac.uk

Vaclav Brezina
Centre for Corpus Approaches to Social Science (CASS)
Faculty of Arts and Social Sciences
Lancaster University
Lancaster, LA1 4YD
UK
v.brezina@lancaster.ac.uk