The Spoken BNC2014

Designing and building a spoken corpus of everyday conversations

Robbie Love[i], Claire Dembry[ii], Andrew Hardie[i], Vaclav Brezina[i] and Tony McEnery[i]
[i]Lancaster University / [ii]Cambridge University Press

This paper introduces the Spoken British National Corpus 2014, an 11.5-million-word corpus of orthographically transcribed conversations among L1 speakers of British English from across the UK, recorded in the years 2012–2016. After showing that a survey of the recent history of corpora of spoken British English justifies the compilation of this new corpus, we describe the main stages of the Spoken BNC2014's creation: design, data and metadata collection, transcription, XML encoding, and annotation. In doing so we aim to (i) encourage users of the corpus to approach the data with sensitivity to the many methodological issues we identified and attempted to overcome while compiling the Spoken BNC2014, and (ii) inform (future) compilers of spoken corpora of the innovations we implemented to attempt to make the construction of corpora representing spontaneous speech in informal contexts more tractable, both logistically and practically, than in the past.

Keywords: Spoken BNC2014, transcription, corpus construction, spoken corpora

1. Introduction

The ESRC Centre for Corpus Approaches to Social Science (CASS)[1] at Lancaster University and Cambridge University Press have compiled a new, publicly-accessible corpus of present-day spoken British English, gathered in informal contexts, known as the Spoken British National Corpus 2014 (Spoken BNC2014). This

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.

International Journal of Corpus Linguistics 22:3 (2017), **319–344**. DOI 10.1075/ijcl.22.3.02lov
ISSN 1384-6655 / E-ISSN 1569-9811

is the first publicly-accessible corpus of its kind since the spoken component of the original British National Corpus, completed in 1994, which, despite its age, is still used as a proxy for present-day English in research today (e.g. Hadikin 2014, Rühlemann & Gries 2015). The new corpus contains data from the years 2012 to 2016. It is publicly available via Lancaster University's *CQPweb* server as of September 2017, with the underlying XML files downloadable from late 2018. It will subsequently form the spoken component of the larger British National Corpus 2014, whose written component is also under development. The main source of recordings for the project was a national participation campaign. In total, the research team has amassed 11.5 million words of transcribed content, featuring 668 speakers in 1,251 recordings.

This paper describes how we designed and built the Spoken BNC2014. Section 2 presents a review of a number of existing corpora of spoken English. We start by introducing our own corpus's predecessor, the spoken component of the original British National Corpus (hereafter Spoken BNC1994), and then consider other notable English corpora that consist wholly or partly of spoken data compiled after 1994. This review demonstrates that no subsequent corpus fulfils the criteria which appear to have led to the Spoken BNC1994's great utility and longevity in linguistic research. In Section 3, we address decisions relating to corpus design and the collection of both the data (audio recordings) and metadata (speaker and recording information). This design necessarily represents a compromise between the ideally representative corpus and the constraints of what is realistically possible. The compromise we adopt is employing an opportunistic approach to data collection rather than attempting to adhere to a predetermined sampling frame, as discussed in detail in Section 3.1. Likewise, the design of the metadata scheme for the corpus represents a compromise between (i) the need for comparability with the Spoken BNC1994 and (ii) the desirability of improving upon some of the decisions made during the compilation of the Spoken BNC1994. Finally, we discuss the participant recruitment method and our decision to use smartphones as recording devices to collect our data. In Section 4, we describe the development of a bespoke transcription scheme for the Spoken BNC2014, as well as decisions made with regards to XML conversion, POS-tagging and lemmatization.

2. Similar existing corpora – why do we need a new one?

A well-known problem afflicting corpus linguistics as a field is its tendency to prioritise written forms of language over spoken forms, in consequence of the much greater difficulty, higher cost and slower speed of collecting transcribed speech:

> A rough guess suggests that the cost of collecting and transcribing in electronic form one million words of naturally occurring speech is at least 10 times higher than the cost of adding another million words of newspaper text.
>
> (Burnard 2002: 6)

Contemporary online access to newspaper material means that this disparity is likely to be even greater today than in 2002. The resulting bias in corpus linguistics towards a "very much written-biased view" (Lüdeling & Kytö 2008: vi) of language is problematic if one takes the view that speech is the primary medium of communication (Čermák 2009: 113), containing linguistic variables that are important for the accurate description of language, and yet inaccessible through the analysis of corpora composed solely of written texts (Adolphs & Carter 2013: 1).

The Spoken BNC1994 is one of the few widely accessible corpora of spoken British English, and is thus heavily relied upon in corpus research on spoken English. However, that the "go-to" dataset is over twenty years old is, as we will show in Section 2.3, a problem for current and future research. Addressing this situation is the main reason for our undertaking the development of a new spoken corpus.

2.1 The Spoken British National Corpus 1994

The compilation of the Spoken BNC2014 is informed largely by the original BNC's spoken component (see Crowdy 1993, 1994, 1995), which is "one of the biggest available corpora of spoken British English" (Nesselhauf & Römer 2007: 297). The goal of the BNC1994's creators was "to make it possible to say something about language in general" (Nesselhauf & Römer 2007: 5). Thus its spoken component was designed to function as a representative sample of spoken British English (Burnard 2007). It was created between 1991 and 1994, and was designed in two parts: the demographically-sampled (DS) part (c. 40%) and the context-governed part (c. 60%) (Aston & Burnard 1998). The Spoken BNC1994DS, as we call it (also known as the 'conversational part', Leech et al. 2001: 2), contains informal, "everyday spontaneous interactions" (Leech et al. 2001: 2), and its contributors (the volunteers who made the recordings of their interactions with other speakers) were "selected by age group, sex, social class and geographic region" (Aston & Burnard 1998: 31). 124 adult contributors made recordings capturing the language of over 1,000 speakers (Aston & Burnard 1998: 32). In total, the Spoken BNC1994DS contains 4.2 million words of transcribed conversation. The context-governed part (also known as the 'task-oriented part', Leech et al. 2001: 2) contains formal encounters from particular institutional settings, which were "categorised by topic and type of interaction" (Aston & Burnard 1998: 31). The Spoken BNC1994DS also incorporates the Bergen Corpus of London Teenage Language (COLT), a half-million-word sample of

spontaneous conversations among teenagers between the ages of 13–17, collected in a variety of boroughs and school districts in London in 1993 (Stenström et al. 2002).

Despite certain weaknesses in design and metadata, which we discuss in Section 2.3, the Spoken BNC1994 has proven a highly productive resource for linguistic research over the last two decades. It has been influential in the areas of grammar (e.g. Rühlemann 2006, Gabrielatos 2013, Smith 2014), sociolinguistics (e.g. McEnery 2005, Säily 2011, Xiao & Tao 2007), conversation analysis (e.g. Rühlemann & Gries 2015), pragmatics (e.g. Wang 2005, Capelle et al. 2015, Hatice 2015), and language teaching (e.g. Alderson 2007, Flowerdew 2009), among others. Part of the reason for the widespread use of the BNC1994 is that it is an open-access corpus; researchers from around the world can access the corpus at zero cost, either by downloading the full text from the Oxford Text Archive (http://ota.ox.ac.uk/desc/2554), or using the online interfaces provided by various institutions including Brigham Young University (*BNC-BYU*, Davies 2004, http://corpus.byu.edu/bnc/) and Lancaster University (*BNCweb*, Hoffmann et al. 2008, http://bncweb.lancs.ac.uk/bncwebSignup). Yet it is undoubtedly the unique access that the Spoken BNC1994 has provided to large scale orthographic transcriptions of spontaneous speech that has been the key to its success. Such resources are needed by linguists but are expensive and time consuming to produce and hence are rarely accessible as openly and easily as is the BNC1994.

2.2 Other British English corpora containing spoken conversational data

Other corpora of spoken British English exist which are similarly conversational and non-specialized in terms of context. Although they have the potential to be just as influential as the Spoken BNC1994DS, they are much harder to access for several reasons. Some have simply not been made available to the public for commercial reasons. The Cambridge and Nottingham Corpus of Discourse in English (CANCODE), for example, forms part of the Cambridge English Corpus, which is a commercial resource belonging to Cambridge University Press and is not accessible to the wider research community (Carter 1998: 55). Other corpora are available only after payment of a license fee, which makes them generally less accessible. For instance, Collins publishers' WordBanks Online (https://www.collins.co.uk/page/Wordbanks+Online) offers paid access to a 57-million-word subcorpus of the Bank of English (containing data from British English and American English sources, 61 million words of which is spoken); the charges range, at time of this writing, from a minimum of £695 up to £3,000 per year of access. Likewise, the British component of the International Corpus of English (ICE-GB), containing one million words of written and spoken data from the 1990s (Nelson et al. 2002: 3), costs over £400 for a single, non-student license (http://www.ucl.ac.uk/english-usage/projects/ice-gb/

iceorder2.htm). Some other corpora are generally available, but sample a more narrowly defined regional variety of English than simply 'British English'. For instance, the Scottish Corpus of Texts and Speech (SCOTS) (Douglas 2003), while it is free to use, contains only Scottish English and no other regional varieties of English from the British Isles.

These restrictions appear to have translated into a much lower level of research output using these datasets. As a crude proxy for the academic impact of these corpora, we searched for publications using them in Lancaster University's online library system. At the time of writing, a search for CANCODE retrieves 54 publications; *WordBanks Online* only 45; ICE-GB 300; and SCOTS 34. By contrast, searching for the BNC1994 identifies 3,000 publications. While an admittedly rough rule of thumb, this quick search shows that even though conversational, non-specialized spoken corpora that may be just as useful as the Spoken BNC1994DS have been compiled since 1994, their limited availability, and/or the expense of accessing them, has meant that the Spoken BNC1994 is almost certainly the most widely used spoken corpus of British English to date.

2.3 Justification for the Spoken BNC2014

It is clearly problematic that research into spoken English is still using a 23-year-old corpus to explore 'present-day' English. The reason why no spoken corpus since the Spoken BNC1994 has equalled its utility for research seems to be that no other corpus has matched all four of its key strengths:

i. orthographically transcribed data
ii. large size
iii. general coverage of spoken British English
iv. (low or no cost) public access

Each of the other projects mentioned above fails to fulfil one or more of these criteria.

The problem of the Spoken BNC1994's continued use would be lessened if it were not still treated as a proxy for present-day English – i.e. if its use were mainly historical – but this is not the case. For researchers interested in spoken English who do not have access to privately held spoken corpora this is unavoidable; the Spoken BNC1994 is still clearly the best publicly-accessible resource for spoken British English for the reasons outlined. Yet as time has passed, the corpus has been used for purposes for which it is becoming increasingly less suitable. For example, a recent study by Hadikin (2014), which investigates the behaviour of articles in spoken Korean English, uses the Spoken BNC1994 as a reference corpus of present-day English. Appropriately, Hadikin (2014) gives the following warning:

> With notably older recordings [than the Korean corpora he compiled] […] one has to be cautious about any language structures that may have changed, or may be changing, in the period since then. (Hadikin 2014: 7)

In this respect, Hadikin's (2014) work typifies a range of recent research which, in the absence of a suitable alternative, uses the Spoken BNC1994 as a sample of present-day English. The dated nature of the Spoken BNC1994 is demonstrated by the presence in the corpus of references to public figures, technology, and television shows that were contemporary in the early 1990s; see Examples (1) to (3):

(1) Oh alright then, so if John Major gets elected then I'll still [unclear][2] (BNC, KCF 105)

(2) Why not just put a video on?[3] (BNC, KBC 18)

(3) Did you see The Generation Game?[4] (BNC, KCT 2546)

We see, then, the need for a new corpus of conversational British English to allow researchers to continue the kinds of research that the Spoken BNC1994 has fostered over the past two decades. This new corpus will also make it possible to turn the ageing of the Spoken BNC1994 into an advantage – if it can be compared to a comparable contemporary corpus, it could become a useful resource for exploring recent change in spoken English. The Spoken BNC2014 project enables scholars to realise these research opportunities as well as, importantly, allowing *gratis* public access to the resulting corpus.

3. Corpus design and data collection

A key decision we made early on in the creation of the Spoken BNC2014 was to collect data which occurred only in informal contexts – i.e. data which would be comparable to the Spoken BNC1994DS. Our rationale for excluding context-governed data is simply that we have noted there exists greater use of, and demand for, conversational data. Researchers who wish to study spoken British English occurring in specific contexts are likely to collect their own, specialized corpora. Moreover, some such specialized corpora have been released publicly by their creators and

2. John Major was Prime Minister of the United Kingdom between 1990 and 1997.

3. The VHS tape cassette, or 'video', was a popular medium for home video consumption in the 1980s and 1990s before the introduction of the DVD in the late 1990s.

4. *The Generation Game* was a popular British television gameshow which was broadcast between 1971 and 2002.

are available to researchers with an interest in the defined context in question; an example is BASE, the British Academic Spoken English Corpus, which contains university lectures and seminars (see Thompson & Nesi 2001). So, researchers with an interest in context-governed English speech already have options open to them. A general corpus of informal speech is, however, harder to collect due to the requirements of size and demographic spread – and therefore, in consequence, much more in demand in the research community.

Another decision we faced was what we should do about the known shortcomings of the Spoken BNC1994DS. Most importantly, certain issues exist in the Spoken BNC1994DS in terms of its speaker metadata; it has been criticised for the "often unhelpful" and inconsistent availability of speaker metadata (Lam 2009: 176). Indeed, Burnard (2002: 7) admits that the classifications used to categorise speakers are sometimes "poorly defined" and "partially or unreliably populated". In Sections 3.1 to 3.3, we describe the steps we took to attempt to improve upon this in the Spoken BNC2014. By instructing contributors to use their own devices (smartphones) to make recordings, rather than supplying recording equipment, we could facilitate an opportunistic approach to data collection which required no training for contributors prior to recording. Furthermore, rather than instructing contributors to provide metadata on behalf of all the speakers they recorded, the contributors gathered metadata directly from each speaker and passed it on to the research team.

3.1 Opportunistic data collection

Every corpus compilation project is, by definition, a sampling project (Biber 1993: 243). The appropriateness of the sample depends on factors including the purpose of the research and the domain within which the data is being collected. In the case of the Spoken BNC1994DS, the demographic categories were divided into two types: "selection criteria" and "descriptive criteria" (Burnard 2002: 6). The selection criteria are the gender, age, socio-economic status and region of the speakers. The descriptive criteria were those which were not controlled during the collection of the data but which were recorded for information; these included the domain and type of speech recorded (Burnard 2002: 6).

The aim of the compilers of the Spoken BNC1994 was to enable research "for a wide variety of linguistic interests" (Wichmann 2008: 189). Part of this aim was to assemble a corpus that was as representative as possible of the language variety under investigation, i.e. "the language production of the population of British English speakers in the United Kingdom" (Crowdy 1993: 259).

Before the Spoken BNC1994 was completed, Crowdy (1993) claimed that:

> *representativeness is achieved* by sampling a spread of language producers in terms of age, gender, social group, and region, and recording their language output over a set period of time. (Crowdy 1993: 257, emphasis added)

However, a compromise was clearly made between what would maximize representativeness and what was possible in practice. As Burnard (2002: 5) points out, "no-one could reasonably claim that the corpus was statistically representative of the whole language", although he is clear that the combination of criteria for selection and description would at least encourage proportionality between, and variability within, the demographic categories (Burnard 2002: 6) – the component categories of each selection criterion were predetermined, and target proportions were assigned for each.

Table 1 shows how each category of the BNC1994's selection criteria was populated, according to the number of words produced by speakers.

Table 1. Proportions of words in Spoken BNC1994 assigned across each of the three selection criteria (adapted from Burnard 2000: 13)

Selection criteria	Category	% 'w-units'
Gender	Male	41.14
	Female	58.47
	Unknown	0.38
Age	0–14	6.30
	15–24	15.71
	25–34	20.16
	35–44	19.96
	45–59	22.76
	60+	15.08
Socio-economic status	AB	32.41
	C1	26.08
	C2	25.69
	DE	14.91
	Unknown	0.88

Particularly underrepresented at the time, it appears, were males, children, and the elderly. Despite creating a sampling frame for the selection criteria, the priority in practice seems to have been to collect as much data as possible and to accept the consequent imbalances in the corpus across the demographic categories. This may sound like a careless strategy, but we argue that this was the only reasonable approach given the costs associated with collecting spoken data. Furthermore, some researchers would go on to craft smaller subcorpora of the data, which were more balanced according to given metadata categories (e.g. BNC 64, Brezina & Meyerhoff

2014). This means that, despite imbalances across the corpus as a whole, it was still possible to analyse demographic categories of equal size if one was willing to work with a smaller data set. Making use of a geological metaphor, the Spoken BNC1994 can be viewed as containing a small "core" of data with evenly balanced demographic categories, and a larger "mantle" of additional data which, when combined with the core, produces a large but not balanced corpus.

For the Spoken BNC2014 we decided to adopt a similar approach to the BNC1994: accepting the data that became available, while monitoring the levels of the demographic categories to be alerted to any imbalances that were severe. This is what we call an 'opportunistic approach' to data collection. If any such "holes" in the data began to appear, we attempted to address these by targeting those specific groups of people – variously through Facebook and Twitter advertisement campaigns, student recruitment campaigns at universities, and press releases which targeted speakers of a particular age, or from a certain geographical region. The resulting data set (see Appendix), which is more than twice the size of the Spoken BNC1994DS, represents an improvement in balance when compared to the Spoken BNC1994 – some categories are well balanced (e.g. northern vs. southern speakers) and some categories are better populated than they would have been had we not monitored the numbers and targeted specific social groups (e.g. elderly speakers). However, there are some peaks and troughs – a major trough being the dearth of speakers from Scotland, Wales and Northern Ireland. We accept this dearth because spoken corpora of English spoken by people from these countries have been collected and made available since the release of the Spoken BNC1994. The previously mentioned SCOTS (Douglas 2003) contains approximately one million words of Scottish English speech – most of which was collected in the 2000s. The Bangor Siarad corpus (Deuchar et al. 2014) contains 450,000 words of bilingual Welsh-English spontaneous speech collected between 2005 and 2008. ICE-Ireland (Kallen & Kirk 2008) comprises approximately 300,000 words of spoken data collected from speakers of Northern Irish English in the mid 1990s to early 2000s. However, as discussed in Section 2.2, no comparable corpus containing "English English" has been made publicly available since the Spoken BNC1994DS; and so we prioritized collecting data for England, as that is where the greatest need lay.

This prioritization of England does mean that the full Spoken BNC2014, though not as imbalanced as the Spoken BNC1994DS, is not a properly balanced corpus if taken as a whole. Yet, as noted, it was no more designed to be so than the Spoken BNC1994 was. Our resolution is to explicitly facilitate the analysis of the Spoken BNC2014 both as a full, unbalanced version (maximising the virtue of size), and also as the "core" on its own (a smaller, balanced subcorpus derived from the whole corpus). The core subcorpus contains an approximately equal number of tokens within each category for each of the following criteria: gender, age, socio-economic

status, and English region. Users of the corpus in Lancaster University's *CQPweb* server are able to move between the entirety of the corpus and the core subcorpus as they wish, so that they can select whichever fits better with the purpose at hand. The core/non-core status of different segments of the corpus is also coded as metadata in the XML-format release of the data.

The alternative, non-opportunistic approach – drawing up a sampling frame and actively seeking out recordings from particular groups of speakers – might well have produced a more representative or balanced corpus, but would, at the very least, have undoubtedly taken much longer to produce.[5] That would have worked against our aim to produce a corpus that can – for a while – be plausibly accepted as a proxy for present-day British English. It would also have been prohibitively time consuming to do this, which with a fixed level of resource available would necessarily lead to the end-result corpus being smaller by perhaps an order of magnitude.

3.2 Recruitment of participants and audio recording

One of the most innovative features of the Spoken BNC2014 is the use of PPSR (public participation in scientific research) for data collection (see Shirk et al. 2012). Anyone who was interested in contributing recordings to the Spoken BNC2014 was directed to a website which described the aims of the project and included a contact form to allow them to register their interest in contributing data. People who registered interest were contacted by the Cambridge team via email with further instructions. Public attention was captured by a series of national media campaigns in 2014 and 2015, as well as through our participation in public engagement events such as the Cambridge University *Festival of Ideas* and the UK Economic and Social Science Research Council's *Festival of Social Sciences*. To incentivize participation, we offered payment of £18 for every hour of recording of a sufficient quality for corpus transcription, and, importantly, submission of all associated consent forms and full speaker metadata. All speakers were required to give informed consent prior to recording, and contributors took responsibility for making recordings and for gathering consent and metadata from all speakers they recorded. We used this opportunity to gather metadata from each individual speaker directly, via the contributors, since no contact was made between the research team and the speakers with whom the contributors chose to converse. To ensure that all information and

5. We found that that certain groups (e.g. NS-SEC groups 6 and 7) were less forthcoming with data than others, despite contributors being paid for providing us with recordings. Therefore, it is not guaranteed that a non-opportunistic approach would produce a more balanced corpus, if some groups of the population are largely unwilling to contribute, even with the offer of payment.

consent was captured, no payments were made to contributors until all metadata, consent forms and related documentation was fully completed for each recording.[6]

Given the general availability of digital audio recorders as a built-in capability of smartphones and other widely used consumer devices, our decision to use smartphones to gather the data meant that all recordings were made digitally. Specifically, contributors were instructed to use the built-in audio recording feature in their smartphones to make recordings. This is in contrast to the Spoken BNC1994, which used analogue recording devices, the recordings from which were subsequently digitised (Crowdy 1994: 15). The use by contributors of their own recording equipment greatly reduced the cost associated with arranging for the recordings, as we did not need to purchase equipment, distribute it, train users to use it and collect it back from them.[7]

3.3 Metadata categories in the Spoken BNC2014

Unlike the Spoken BNC1994, speakers in the Spoken BNC2014 provided their own metadata. This gave us the flexibility to collect a larger set of metadata than was collected in the earlier corpus. The following sections introduce the items of metadata that are recorded for each speaker in the corpus.

3.3.1 *Name*

This was retained only for the purpose of communication between the team at Cambridge and the contributors. All names were converted into unique speaker ID codes to maintain de-identification (the removal or coding of identifiable information for public use, while retaining such information privately, Ribaric et al. 2016) before the transcripts were sent to Lancaster for processing (see Section 4). The term 'de-identification' refers to the same process that has hereforeto often been labelled 'anonymization'.

---

6. All data is stored and analysed in compliance with the UK Data Protection Act 1998.

7. Although there is an increasing desire for spoken corpora that "move beyond text and language as conventionally conceptualised" (Adolphs et al. 2015: 61), our goal was to create a corpus that is comparable to the Spoken BNC1994DS. Therefore, we made no attempt to record conversations as video rather than just audio, or to record any other live contextual data – for example the GPS position of the smartphones (cf. Adolphs et al. 2015). Video recording equipment that does not heavily compromise the unobtrusiveness of the recording event (and, therefore, the likelihood of the data being as natural as possible) has yet to become available at low expense (Adolphs & Carter 2013: 147), and we did not have the time to develop a bespoke smartphone application for recording any other data.

3.3.2 *Age*

For most of the speakers in the Spoken BNC2014 the exact age is available as free-form speaker metadata (e.g. "27"). In addition, we have categorised speakers according to two age-based schemes. The first is the Spoken BNC1994 age categorisation scheme: 0–14, 15–24, 25–34, 35–44, 45–59, 60+, and Unknown. This facilitates comparison between the two corpora. The second scheme is more fine-grained: 0–10, 11–18, 19–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99, and Unknown. We increased the number of categories to facilitate more sophisticated apparent-time analysis of the new data; the revised scheme starts with a primary division at 18/19 (18 being the latest age of school-leaving in the UK) and then subdivides the resulting juvenile/adult sections into decades (as closely as possible).

3.3.3 *Gender*

On the consent forms, speakers specified their gender via a free-text box (i.e. they could write whatever they liked in their own words). All speakers self-reported their gender as either "female" or "male", which we code as F or M respectively; a third classification, "n/a (multiple)", is used only for groups of multiple speakers (e.g. in attributing vocalisations such as laughter when produced by several speakers at once).

3.3.4 *Accent/dialect*

Speakers used a free-text box to enter a description of their own dialect (e.g. "Geordie", "Northern", etc.). The self-reported dialect of speakers has then been coded according to a four-level classification scheme (Table 2), the fourth level of which is drawn from the UK government's *Nomenclature of Territorial Units for Statistics* (NUTS, http://webarchive.nationalarchives.gov.uk/20160105160709/ http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/eurostat/index.html). The scheme therefore does not arise from considerations of linguistic classifications of the UK (cf. Trudgill 1999) but rather geopolitical ones. This choice of scheme reflects our principle that the pre-selection of categories for sociolinguistic analysis should not impose assumptions of linguistic patterns upon the corpus but ought rather to allow the data to reveal such patterns. While it might have been preferable for us to develop a categorization and then train the speakers to use it, this would clearly have been infeasibly time-consuming. But self-reported dialect data is not without its own virtues: it is, for instance, of great value to researchers interested in perceptual dialectology (e.g. Montgomery 2012). Moreover, it will be possible (in principle at least) to assign regional-dialect classifications to the recorded speakers according to objective, linguistic criteria at some later point. But it is generally *not* possible to facilitate perceptual dialectology research other than by asking the speakers what variety of English they believe they speak. So, while one driver for our decision to gather self-report data on dialect was practical, another was principled – we wanted to gather from the speakers that information that could not be easily inferred, or

inferred at all, from their data at a later date: the variety of British English they believed themselves to speak.

Table 2. Classification scheme for speaker dialect in the Spoken BNC2014

(1) Global	(2) Country	(3) Supra-region	(4) Region
UK	English	North	North East
			Yorkshire & Humberside
			North West (not Merseyside)
			Merseyside
		Midlands	East Midlands
			West Midlands
		South	Eastern
			South West
			South East (not London)
			London
	Scottish	Scottish	Scottish
	Welsh	Welsh	Welsh
	Northern Irish	Northern Irish	Northern Irish
Non-UK	Irish	Irish	Irish
	Non-UK	Non-UK	Non-UK
Unspecified	Unspecified	Unspecified	Unspecified

Based on speakers' free-text answers to the question of what variety of English they speak, each speaker is assigned to a category in each of the four levels in Table 2 ("global", "country", "supraregion" and "region"). The assignments depend upon how much could be inferred from their self-reported response. Our aim was to maximise specificity (in other words, to "get as much out of" the metadata as possible while allowing speakers to describe themselves in their own words). For example, a speaker who entered "Geordie" would be assigned to: (Level 1 – UK; Level 2 – English; Level 3 – North; Level 4 – North East). A speaker who entered "Northern" would be assigned to: (Level 1 – UK; Level 2 – English; Level 3 – North; Level 4 – Unspecified). Thus, a level 4 analysis would exclude a self-reported "northern" speaker and place them in the "unspecified" category because the specific region of the north to which they refer (if any) is not known. It should also be noted that analysing the data at the third level ("supra-region") facilitates comparison with the regional classification in the Spoken BNC1994 – albeit the latter is itself not unproblematic.

3.3.5 *Occupation*

Speaker occupation is coded according to the *National Statistics Socio-economic Classification* (NS-SEC), which has been the government standard for the UK census since 2001, and is also used in the UK Labour Force Survey. To complement this scheme, the UK government makes available a free online tool (https://onsdigital.

github.io/dp-classification-tools/standard-occupational-classification/ONS_SOC_occupation_coding_tool.html) which converts the name of any given occupation into an NS-SEC code. NS-SEC is of course a different scheme to that used in the Spoken BNC1994. That scheme, *Social Grade*, was never used by the government and instead was (and is) popular in the market research sector. No formal standard has been established for translating either of these schemes to the other, but in the interests of comparability we have proposed an automatic mapping from NS-SEC to Social Grade so that both schemes can be analysed in the Spoken BNC2014 (Table 3).

Table 3. Mapping between the NS-SEC and Social Grade assumed for Spoken BNC2014 speaker metadata

NS-SEC	Description		Social Grade	Description
1	Higher managerial, administrative and professional occupations:[8]		A	Higher managerial, administrative and professional
1.1	Large employers and higher managerial and administrative occupations			
1.2	Higher professional occupations			
2	Lower managerial, administrative and professional occupations	MAPS ON TO …	B	Intermediate managerial, administrative and professional
3	Intermediate occupations		C1	Supervisory, clerical and junior managerial, administrative and professional
4	Small employers and own account workers			
5	Lower supervisory and technical occupations		C2	Skilled manual workers
6	Semi-routine occupations		D	Semi-skilled and unskilled manual workers
7	Routine occupations			
8	Never worked and long-term unemployed		E	State pensioners, casual and lowest grade workers, unemployed with state benefits only
*	Students/unclassifiable			

8. This is not in and of itself an analytic category; rather it comprises analytic categories 1.1 and 1.2

3.3.6 *Other metadata categories*

The other speaker metadata categories we collected were:

i. Nationality
ii. Birthplace
iii. Current location
iv. Duration of stay there
v. Mother tongue
vi. Most influential country on language
vii. Additional languages
viii. Education level

The metadata categories pertaining to the recordings were:

i. Number of speakers
ii. Recording location
iii. Conversations topics (assigned by contributor)
iv. Main conversation topic (assigned by contributor)

These items of speaker and recording metadata are entirely self-reported; the form in which the speakers provided this information is reproduced verbatim in the corpus metadata and documentation without attempts to schematize or standardize.

4. Transcribing the Spoken BNC2014

Transcription of the Spoken BNC2014 recordings was undertaken by a team of transcribers employed by Cambridge University Press. They were trained according to a bespoke transcription scheme developed for this project (see Section 4.1). In Sections 4.1 to 4.3, we discuss the decisions we made about the transcription of the Spoken BNC2014 recordings, as well as the steps taken to convert the resulting transcripts into a suitable XML-based canonical format for distribution and archiving of the corpus.

4.1 Developing the transcription scheme

The first two questions that Crowdy (1994) poses in his account of the Spoken BNC1994's transcription system are "who is the transcription for?" and "how will it be used?" (Crowdy 1994: 25), foregrounding the importance of purpose when transcribing spoken data. Our starting point was to employ a level of "standard orthographic" transcription (Leech et al. 2001: 12) that was simple and easy to

implement. Like the Spoken BNC1994, the main aim of the Spoken BNC2014 is to facilitate the quantitative study of "morphology, lexis, syntax, pragmatics, etc." (Atkins et al. 1992: 10), allowing users to search for "particular features or patterns" and view them "in concordanced form" (Crowdy 1995: 228). An orthographic transcription serves the needs of research in these areas. We explicitly exclude the study of phonetics (segmental or prosodic) from the list of areas that the corpus caters for. While we are of course entirely open to phoneticians making use of the corpus if they wish and so far as they can, most phonetic research typically requires both (i) access to high-quality audio recordings and (ii) full phonetic transcription, neither of which was a possibility within the constraints of this project. Phonetics aside, transcribing recordings in the form of an "idealized 'script' (like a screenplay or drama script) is sufficient for a wide variety of linguistic studies" (Atkins et al. 1992: 10).

The next decision to be made related to the precise nature of the scheme for orthographic transcription to be employed. This can be a highly consequential matter, as it affects the time taken for transcription, and thus the cost and therefore the possible size of the corpus (cf. Schmidt 2016). The key requirement is for a robust transcription scheme that, critically, minimizes the level of transcriber inference that is needed – that is, the number of decisions that a transcriber must make about potentially ambiguous speech phenomena. Speech phenomena which require a higher level of transcriber inference to be included in linguistic detail, such as "false starts, hesitation, non-verbal signals" (Atkins et al. 1992: 10), take more time to transcribe, and even more time to achieve consistency within each transcriber's work and across transcribers. We aimed, therefore, to normalize or disregard these phenomena at the transcription stage as far as we could, while still serving most of the needs of most of our intended users.

Defining such a robust scheme meant that all of the issues likely to be encountered by transcribers had to be explored, and decisions made about how to deal with them, before full scale transcription commenced. We found that Atkins et al. (1992: 11–12) provide a good starting point in terms of general recommendations for approaching spoken corpus transcription. These recommendations include: beginning each turn with an identifying encoding of the speaker; marking inaudible segments; normalizing numbers and abbreviations; and producing a "closed set of permissible forms" for the transcription of dialect and non-standard words. Atkins et al. (1992) also advise careful thought about the extent to which punctuation should represent written conventions, and suggest that faithful and precise transcription of overlapping speech is costly; thus, an evaluation of the value and utility of including both punctuation and overlaps should be made before transcription begins.

Similarly, with regard to functional and non-functional sounds (also known as filled pauses, or more informally *um*s and *ah*s), Atkins et al. (1992) note that classifying these speech sounds according to discourse function requires a high level

of inference on the part of the transcriber. Therefore "a large set of orthographic representations" (Atkins et al. 1992: 12) of speech sounds, rather than their possible functional mappings, should be added to the transcription scheme. That is, transcribers should be instructed to select a transcription for each *um* or *ah* based only on its sound form, and should not attempt to imbue meaning into the transcription of these non-lexical sounds (e.g. by providing pragmatic annotation). Rather, as Atkins et al. (1992) suggest, the interpretation of such sounds should be left to researchers with access to the recordings who choose to investigate these phenomena at a later date. This recommendation can be seen as a specific case of Crowdy's (1994: 25) more general principle that researchers should use the transcript as a "baseline" and that analysis beyond the scope of a simple orthographic transcription should be undertaken by those researchers who wish to "analyse the text in more detail".

Admittedly, such additional analysis will not immediately be possible on release of the corpus, because we have not been able to de-identify the audio recordings from the Spoken BNC2014 within the scope of the present project (de-identification being necessary to preserve speakers' privacy). However, we will in future pursue further funding to de-identify and release the original recordings, thus enabling functional analysis of speech features currently transcribed without any pragmatic/discourse classification. The benefit of an approach which omits any features requiring inferential decisions by the transcribers is not merely theoretical; rather, we have practical evidence of its usefulness. During the pilot phase of our work, we undertook an experiment in which we asked the transcribers to annotate any segment of an utterance containing reported direct speech (that is, material that the speaker is quoting from elsewhere) during their transcription of the audio. The transcribers reported that this task was not difficult. However, when their work was compared to a standardised transcript analysed by a linguist, they were found to have marked less than a third of qualifying clauses. We thus see that requiring transcribers to include detailed analytic distinctions either leads to low quality results, or necessitates a high level of *post hoc* correction by a linguist. Neither of these outcomes was desirable or affordable. We were, therefore, convinced of the need to make the transcription scheme exclude not only the annotation of quoted speech but also any other type of additional annotation that would require the input of a linguist – though, as noted, such additions to the basic transcription will of course be welcome after the release of the audio data.

Given the above points, our first decision was to avoid simply re-using the Spoken BNC1994 transcription scheme, as documented by Crowdy (1994). The reason for this is that Crowdy's (1994) account of the Spoken BNC1994 transcription conventions is by no means comprehensive; only sixteen features are identified and the entire scheme is less than two thousand words in length. Furthermore, not enough examples are provided to eliminate ambiguity, and some of the examples

which are provided are transcribed inconsistently. For example, full stops and commas are to be used to mark "a syntactically appropriate termination or pause in an utterance, approximating to use in written text", and an ellipsis to mark "a longer pause – up to 5 seconds" (Crowdy 1994: 27). But in practice, the examples include uses of full stops and commas in positions that *would not* license a punctuation mark in written English, as shown in Example (4), suggesting that the full stop/comma versus ellipsis rule was not followed by transcribers in a consistent manner:

(4) <2> I think it's always, deceptive on days like this because its, overcast and [er] […]
<2> But, but er, he's … just broken away from his girlfriend and [<unclear>]
<1> [Oh has] he, oh. Well he seemed happy enough when he called.
(Crowdy 1994: 28)

Furthermore, Crowdy's (1994) scheme states that question marks are to be used to indicate "questioning" utterances, but this is not done consistently in the examples provided, as in Example (5):

(5) <1> It's a funny old day isn't it.
<2> Mm it's not cold is it? (Crowdy 1994: 28)

It thus seemed appropriate not to apply the 1994 scheme again without thorough review. This is not to imply that the original scheme is wrong; many of the recommendations, we believe, are sensible. However, considering examples such as those above, we were concerned that the transcription scheme as it was did not give enough detail about enough features to maximally ensure inter-transcriber consistency. So, instead:

i. we conducted a critical evaluation of Crowdy's (1994) scheme, identifying which features should be retained, abandoned or adapted;
ii. we reviewed evidence from other work on spoken corpus transcription published since the Spoken BNC1994's compilation, with a particular focus on spoken components of the Cambridge English Corpus as well as recent work at Lancaster University on spoken corpora;
iii. we conducted a small pilot study (as mentioned above) to test some of the proposed features in practice.

The resulting transcription scheme can be found in the user documentation (Love et al. 2017) for the Spoken BNC2014 (which will be available to read online at http://corpora.lancs.ac.uk/bnc2014 as well as included within the corpus distribution download). After each recording was transcribed, the transcript was put though two stages of checking at Cambridge University Press – audio-checking and proofreading – before being sent to Lancaster for processing. At the audio-checking stage, a 5% sample of the recording was checked against the transcript for linguistic

accuracy. If errors were found, the entire recording was checked. After this, the transcript was proofread for errors with regard to the transcription conventions (without reference to the audio). Despite this checking, complete accuracy of transcription cannot, of course, be assumed – even though the scheme has been limited to a basic, orthographic level of transcription. It is unavoidable that the involvement of over a dozen human transcribers (as was the case in the production of the Spoken BNC2014) will lead to certain inconsistencies of transcription decisions. Our extended and elaborated transcription scheme enabled us to minimize – but we would not claim to eradicate – such inconsistency. Indeed, it would be naïve to assume the latter. For example, let us consider the variant pronunciations of the tag question *isn't it*, as represented orthographically by *isn't it, ain't it, innit*, etc. The transcription scheme lists these as permissible non-standard forms, and ideally we would therefore expect each instance of the tag question to have been faithfully transcribed using the spelling variant that matches the actual pronunciation. But in practice, it is very unlikely that a match between non-standard orthography and precise phonetic quality was achieved consistently, both within the transcripts of a given transcriber and indeed between transcribers. As such, we encourage users to consider the data not as a definitive representation of the original speech event, but rather to bear in mind that the transcriptions have been produced under the constraints of what we now believe to be the natural, terminal limit of consistency between human transcribers.

4.2 Speaker identification

None of the previous work that we consulted when developing the transcription scheme had recognised one important consideration: the degree of confidence with which transcribers were able to identify the speaker responsible for each turn (especially in recordings which contain more than two speakers). Evidence from a pilot study on this issue suggests that transcribers tend to assign their "best guess" speaker to a given turn – resulting in inaccurate speaker identification in cases where they guess incorrectly. To account for this, we introduced a new speaker ID convention which allowed transcribers to indicate when they were not fully certain in their choice of ID code. The purpose of this is to caution users of the corpus against blindly assuming that all of the speaker ID codes in the corpus texts have been assigned accurately. In Example (6), for instance, the transcriber has indicated that they were not fully certain of which speaker produced the second turn, but that their best guess is speaker S0514 (the *[??]* indicator of low confidence shown here represents an underlying XML attribute-value pair; see below).

(6) S0511: well what happens in the sessions?
S0514[??]: there was some watching videos and stuff (BNCBB001)

Though this measure does not actually improve the accuracy of speaker identification, it does promote user awareness of potential issues with it. Furthermore, this utterance-level attribute data makes it possible to restrict corpus queries to exclude those turns with low confidence in speaker identification. In total, 29,369 utterances (2.45% of utterances; 170,806 tokens) fall into the low confidence category. In addition, we observed that transcriber confidence decreased as the number of speakers in the corpus texts increased, and that texts containing four speakers or more (294 texts; 23.5%) are liable to high degrees (average 40%) of inaccuracy in speaker identification. We therefore recommend that users who require speakers to be attributed accurately use the restricted query function in *CQPweb* (or, equivalently, appropriate preprocessing of a downloaded copy of the XML-formatted corpus) to exclude texts containing four or more speakers.

4.3 Converting the transcripts

The two established standard formats for corpus data interchange and archiving are (i) plain text and (ii) plain text enhanced with markup using XML. Transcripts of spoken data almost always include features beyond the actual words of the text (e.g. indicators of utterance boundaries) and thus XML is the appropriate choice of format. However, XML is a somewhat cumbersome format for direct data entry, and is also rather difficult to teach to non-specialist audio transcribers. It is also challenging to check for accuracy by eye. From the outset, then, we knew that while our goal was to release a canonical version of the corpus in XML, this would not be the system used for transcription. Instead, we designed the transcription scheme to be human-friendly, while making sure that all of its elements could be unambiguously mapped to XML at a later stage. For that reason, the original transcripts used short, easy-to-type codes for features such as utterance boundaries, speaker labels, vocalisations, and de-identified elements (names of people, places, and other potentially identifying information). As the recordings were transcribed by the Cambridge team, the transcripts were sent in batches to the Lancaster team, who used a set of automated conversion scripts to translate the transcripts into XML – at the same time applying a series of further automatic checks on the correct use of the transcription conventions that were not possible prior to conversion to a structured document. This approach was by no means an innovation – the transcription scheme presented by Crowdy (1994) for the Spoken BNC1994 was likewise converted to SGML (and, later, XML) in the released BNC1994.

In consequence, while (as previously noted) we do include a complete description of the human-friendly conventions used in transcription within the corpus documentation (Love et al. 2017), these conventions are not used in the text of the corpus itself; instead, the actual corpus text contains the canonical XML. The

transcription scheme is, then, part of our record of how the corpus was created. It is not exclusively a guide for users. We make it available to users of the corpus in order to make the decisions discussed above absolutely transparent, but also in the hope that it may prove useful as a point of departure for other researchers working on the creation of spoken corpora of this kind.

A number of systems for the use of XML in corpus encoding have been proposed as standards. These include the Text Encoding Initiative (TEI; see Burnard & Bauman 2013) and the Corpus Encoding Standard (CES; see Ide 1996). The former of these was used for (and, in fact, originally developed alongside) the BNC1994. However, as argued by Hardie (2014), these standards are fairly top-heavy and require much more extensive and detailed XML markup than is either necessary or useful for the vast majority of corpus linguistic research. For that reason, rather than use TEI, we opted to follow the recommendations of Hardie (2014) for a "modest" level of XML. We made use of the XML tags and attributes noted by Hardie (2014: 94–101) as having become more-or-less established as *de facto* standard – most of which are in fact also part of TEI and CES; we made additions to this set of codes only where our transcription system required it. For instance, utterances are marked up with <*u*> tags, and each utterance has a *who* attribute, containing the unique ID code of the speaker. These are exactly as described by Hardie (2014) and originate in TEI. However, we also added a *whoConfidence* attribute, which records the transcriber's level of confidence in the speaker attribution (as per the discussion above). The text headers in the corpus use a drastically simpler (and more flatly organised) set of metadata tags than TEI/XML, each element being generated automatically, on a mostly one-to-one basis, from some column of the metadata tables originally collected alongside the recordings. Both the header and body tags are listed in full in the corpus documentation, which also includes a full Document Type Definition (DTD) covering all elements and attributes.

The virtues of making standard types of analytic annotation available to all users of a corpus, by distributing a tagged version alongside the untagged text, are well understood. In line with this principle, we have tagged the whole corpus for part-of-speech (POS) and lemma using the same systems as the original BNC1994 – most notably the CLAWS tagger (see Garside & Smith 1997). However, in a departure from the practice of the BNC1994, we use the C6 tagset instead of the simpler C5 tagset.[9] C5 tags were used in order to achieve a simpler (and thus more reliable) system of POS tagging in the first release of the BNC. However, later BNC1994 releases use a parallel system of simple tags, or major wordclasses, alongside the C5 tags. This system uses one single tag for all nouns, another single tag for all verbs, and so on, and, in our view, addresses the need for a lower-complexity grammatical

9. Both tagsets are available on the CLAWS website: http://ucrel.lancs.ac.uk/claws/

classification highly effectively. Thus, the combination of full-complexity C6 annotation and low-complexity simple tags is the best way to address all the purposes covered by the mid-complexity C5 tags. In the canonical form of the data, all three annotations (C6 POS tags, simple POS tags, and lemmas) are coded as XML attributes on the <*w*> element (which encloses each word: used only in the tagged version of the corpus).

However, as noted above, while the XML release will represent the canonical form of the corpus, the initial release was via Lancaster University's *CQPweb* server. *CQPweb* is the online interface component of the Corpus Workbench software (see Hardie 2012). *CQPweb* provides full support for a number of features which use of the Spoken BNC2014 requires, namely (i) access to all layers of corpus annotation; (ii) restricting analyses to utterances whose speakers fulfil certain demographic criteria of dialect, age, or gender; and (iii) limiting access only to users who have signed the corpus licence. XML elements encoded within a *CQPweb* corpus can be used to control the appearance of the text in concordance lines and other aspects of the interface; on the Lancaster server, we configure the system to display utterance boundaries and speaker ID codes in an easily readable format. So, for instance, the underlying XML attribute-value pair *whoConfidence="low"* – which appears on the <*u*> element – is rendered in the interface as *[??]*, the notation shown in Example (6) above. The display format that we use for such features in *CQPweb* does not replicate the original codes as typed by the transcribers; the display codes were instead devised afresh for maximal visual distinctiveness.

5. Conclusion

In this paper, we have presented a general overview of the design and compilation process of the Spoken BNC2014. Our aim has been to make clear the most important decisions we have made as we have collected, transcribed and processed the data. The resulting corpus (and, eventually, its written counterpart) should be of use to many researchers, educators and students in the corpus linguistics and English language community and beyond. In the short term, we are pleased to note the research presented in this special issue of IJCL, all of which uses the Spoken BNC2014 Sample (see this issue's Editorial), and we anticipate the publication of ground breaking sociolinguistic research based upon this data in the forthcoming year (Brezina et al. forthcoming). Furthermore, CASS has started a new project addressing the creation of a balanced sociolinguistic core form both the Spoken BNC2014 and the BNC1994DS (Brezina et al. 2016). The project combines the expertise from the fields of corpus linguistics and variationist sociolinguistics to develop subsamples of the two larger corpora that will allow sophisticated sociolinguistic searches and analyses.

References

Adolphs, S., & Carter, R. (2013). *Spoken Corpus Linguistics: From Monomodal to Multimodal*. Abingdon: Routledge.

Adolphs, S., Knight, D., & Carter, R. (2015). Beyond modal spoken corpora: A dynamic approach to tracking language in context. In P. Baker & T. McEnery (Eds.), *Corpora and Discourse Studies: Integrating Discourse and Corpora* (pp. 41–62). Houndsmill: Palgrave Macmillan. doi:10.1057/9781137431738_3

Alderson, C. J. (2007). Judging the frequency of English words. *Applied Linguistics*, 28(3), 383–409. doi:10.1093/applin/amm024

Aston, G., & Burnard, L. (1998). *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Atkins, A., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, 7(1), 1–16. doi:10.1093/llc/7.1.1

Biber, D. (1993). Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4), 243–257. doi:10.1093/llc/8.4.243

Brezina, V., & Meyerhoff, M. (2014). Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics*, 19(1), 1–28. doi:10.1075/ijcl.19.1.01bre

Brezina, V., Gablasova, D., McEnery, T., & Meyerhoff, M. (2016). *British National Corpus (BNC) as a sociolinguistic dataset: Exploring individual and social variation*. Retrieved from http://gtr.rcuk.ac.uk/projects?ref=ES%2FP001599%2F1 (last accessed November 2016).

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.

Burnard, L. (2000). Reference Guide for the British National Corpus (World Edition). *Oxford University*. Retrieved from http://www.natcorp.ox.ac.uk/archive/worldURG/urg.pdf (last accessed December 2013).

Burnard, L. (2002). Where did we go wrong? A retrospective look at the British National Corpus. In B. Kettemann & G. Markus (Eds.), *Teaching and Learning by Doing Corpus Analysis* (pp. 51–71). Amsterdam: Rodopi. doi:10.1163/9789004334236_007

Burnard, L. (2007). Reference guide for the British National Corpus (XML Edition). Oxford University. Retrieved from http://www.natcorp.ox.ac.uk/docs/URG/ (last accessed December 2013).

Burnard, L., & Bauman, S. (Eds.) (2013). TEI: P5 Guidelines. *TEI Consortium*. Retrieved from http://www.tei-c.org/Guidelines/P5/ (last accessed June 2017).

Carter, R. (1998). Orders of reality: CANCODE, communication, and culture. *ELT Journal*, 52(1), 43–56. doi:10.1093/elt/52.1.43

Cappelle, B., Dugas, E., & Tobin, V. (2015). An afterthought on let alone. *Journal of Pragmatics*, 80, 70–85. doi:10.1016/j.pragma.2015.02.005

Čermák, F. (2009). Spoken corpora design: Their constitutive parameters. *International Journal of Corpus Linguistics*, 14(1), 113–123. doi:10.1075/ijcl.14.1.07cer

Crowdy, S. (1993). Spoken corpus design. *Literary and Linguistic Computing*, 8(4), 259–265. doi:10.1093/llc/8.4.259

Crowdy, S. (1994). Spoken corpus transcription. *Literary and Linguistic Computing*, 9(1), 25–28. doi:10.1093/llc/9.1.25

Crowdy, S. (1995). The BNC spoken corpus. In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-Up and Annotation* (pp. 224–234). Harlow: Longman.

Davies, M. (2004). BYU-BNC (Based on the British National Corpus from Oxford University Press). *Brigham Young University*. Retrieved from http://corpus.byu.edu/bnc/ (last accessed June 2017).

Deuchar, M., Davies P., Herring J., Parafita Couto, M., & Carter D. (2014). Building bilingual corpora. In E. M. Thomas & I. Mennen (Eds.), *Advances in the Study of Bilingualism* (pp. 93–111). Bristol: Multilingual Matters.

Douglas, F. (2003). The Scottish Corpus of Texts and Speech: Problems of corpus design. *Literary and Linguistic Computing*, 18(1), 23–37. doi:10.1093/llc/18.1.23

Flowerdew, J. (2009). Corpora in language teaching. In M. H. Long & C. J. Doughty (Eds.), *The Handbook of Language Teaching* (pp. 327–350). Oxford: Wiley-Blackwell. doi:10.1002/9781444315783.ch19

Gabrielatos, C. (2013). If-conditionals in ICLE and the BNC: A success story for teaching or learning? In S. Granger, G. Gilquin & F. Meunier (Eds.), *Twenty Years of Learner Corpus Research: Looking back, Moving ahead* (pp. 155–156). Louvain-la-Neuve: Presses Universitaires de Louvain.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech & A. McEnery (Eds.), *Corpus Annotation: Linguistic Information from Computer Text Corpora* (pp. 102–121). London: Longman.

Hadikin, G. (2014). *A*, *an* and *the* environments in Spoken Korean English. *Corpora*, 9(1), 1–28. doi:10.3366/cor.2014.0049

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

Hardie, A. (2014). Modest XML for Corpora: Not a standard, but a suggestion. *ICAME Journal* 38, 73–103. doi:10.2478/icame-2014-0004

Hatice, C. (2015). *Impoliteness in Corpora: A Comparative Analysis of British English and Spoken Turkish*. Sheffield: Equinox.

Hoffmann, S., Evert, S., Lee, D., & Ylva, B. (2008). *Corpus Linguistics with BNCweb: A Practical Guide*. Frankfurt am Main: Peter Lang.

Ide, N. (1996). Corpus Encoding Standard. *Expert Advisory Group on Language Engineering Standards (EAGLES)*. Retrieved from http://www.cs.vassar.edu/CES/ (last accessed June 2017).

Kallen, J. L., & Kirk, J. (2008). *ICE-Ireland: A User's Guide Documentation to accompany the Ireland Component of the International Corpus of English (ICE-Ireland)*. Belfast: Cló Ollscoil na Banríona. Retrieved from http://www.johnmkirk.co.uk/johnmkirk/documents/003647.pdf (last accessed June 2017).

Lam, P. (2009). The making of a BNC customised spoken corpus for comparative purposes. *Corpora*, 4(1), 167–188. doi:10.3366/E174950320900029X

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Harlow: Pearson Education Limited.

Lüdeling, A., & Kytö, M. (2008). Introduction. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. i–xii). Berlin: Walter de Gruyter. doi:10.1515/9783110211429

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.

McEnery, T. (2005). *Swearing in English: Bad Language, Purity and Power from 1586 to the Present*. New York, NY: Routledge.

Montgomery, C. (2012). The effect of proximity in perceptual dialectology. *Journal of Sociolinguistics*, 16(5), 638–668. doi:10.1111/josl.12003

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam/Philadelphia: John Benjamins. doi:10.1075/veaw.g29

Nesselhauf, N., & Römer, U. (2007). Lexical-grammatical patterns in spoken English: The case of the progressive with future time reference. *International Journal of Corpus Linguistics*, 12(3), 297–333. doi:10.1075/ijcl.12.3.02nes

Ribaric, S., Ariyaeeinia, A., & Pavesic, N. (2016). De-identification for privacy protection in multimedia content: A survey. *Signal Processing: Image Communication*, 47, 131–151.

Rühlemann, C. (2006). Coming to terms with conversational grammar: 'Dislocation' and 'dysfluency'. *International Journal of Corpus Linguistics*, 11(4), 385–409. doi:10.1075/ijcl.11.4.03ruh

Rühlemann, C., & Gries, S. (2015). Turn order and turn distribution in multi-party storytelling. *Journal of Pragmatics*, 87, 171–191. doi:10.1016/j.pragma.2015.08.003

Säily, T. (2011). Variation in morphological productivity in the BNC: Sociolinguistic and methodological considerations. *Corpus Linguistics and Linguistic Theory*, 7(1), 119–141. doi:10.1515/cllt.2011.006

Schmidt, T. (2016). Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. *International Journal of Corpus Linguistics*, 21(3), 396–418. doi:10.1075/ijcl.21.3.05sch

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

Smith, A. (2014). Newly emerging subordinators in spoken/written English. *Australian Journal of Linguistics*, 34(1), 118–138. doi:10.1080/07268602.2014.875458

Stenström, A. -B., Andersen, G., & Hasund, I. K. (2002). *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings*. Amsterdam/Philadelphia: John Benjamins. doi:10.1075/scl.8

Thompson, P., & Nesi, H. (2001). The British Academic Spoken English (BASE) Corpus Project. *Language Teaching Research*, 5(3), 263–264.

Trudgill, P. (1999). *The Dialects of England* (2nd ed.). Oxford: Blackwell Publishing Ltd.

Wang, S. (2005). Corpus-based approaches and discourse analysis in relation to reduplication and repetition. *Journal of Pragmatics*, 34(4), 505–540. doi:10.1016/j.pragma.2004.08.002

Wichmann, A. (2008). Speech corpora and spoken corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook* (pp. 187–206). Berlin: Walter de Gruyter.

Xiao, R., & Tao, H. (2007). A corpus-based sociolinguistic study of amplifiers in British English. *Sociolinguistic Studies*, 1(2), 231–273. doi:10.1558/sols.v1i2.241

Authors' addresses

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

Claire Dembry
Cambridge University Press, University Printing House
Shaftesbury Road
Cambridge, CB2 8BS
United Kingdom
cdembry@cambridge.org

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

Vaclav Brezina
ESRC 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

Tony McEnery
ESRC 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