CORPUS DESCRIPTION

Functional variation in the Spoken BNC2014 and the potential for register analysis

Robbie Love[1], Vaclav Brezina[2], Tony McEnery[2], Abi Hawtin[2], Andrew Hardie[2] & Claire Dembry[3]
[1] University of Leeds, UK | [2] Lancaster University, UK |
[3] Cambridge University Press, UK

This article focuses on how register considerations informed and guided the design of the spoken component of the British National Corpus 2014 (Spoken BNC2014). It discusses why the compilers of the corpus sought to gather recordings from just one broad spoken register – 'informal conversation' – and how this and other design decisions afforded contributors to the corpus much freedom with regards to the selection of situational contexts for the recordings. This freedom resulted in a high level of diversity in the corpus for situational parameters such as *recording location* and *activity type*, each of which was captured in the corpus metadata. Focussing on these parameters, this article provides evidence for functional variation among the texts in the corpus and suggests that differences such as those observed presently could be analysable within the existing frameworks for analysis of register variation in spoken and written language, such as multi-dimensional analysis.

Keywords: corpora, corpus design, BNC2014, British English, spoken language

1. Introduction

The Spoken BNC2014 is one of two components of the new British National Corpus 2014, a large dataset representing current British English used in different situations. This article explores, and considers the implications of, the design decisions taken in the process of creating the Spoken BNC2014 within the context of the overall goal of building a large representative corpus of current British English (the BNC2014). Register, defined as "a variety associated with a particular

https://doi.org/10.1075/rs.18013.lov
Register Studies 1:2 (2019), pp. 296–317. issn 2542-9477 | e-issn 2542-9485

situation of use" (Biber & Conrad 2009:6), can be seen as one of the organising principles in the compilation of general corpora, which are designed to represent a wide range of uses of language across different situations. The concept of register (further discussed in Section 2) can be applied both to macro-situations (when contrasting, for example, speech and writing) and to micro-level situations (contrasts among dinner table conversations, museum visits, walks, parties, and so on); so functional register analysis is only desirable for micro-level situations.

This article briefly outlines certain key principles related to register representation in the BNC2014 (Section 2). It goes on to lay out the reasoning that underlies the design decisions made when creating the Spoken BNC2014 (Section 3), as well as the implications of these design decisions for the ultimate register structure of the Spoken BNC2014 (Section 4).

2. BNC2014: Approaches to register

The compilation of the BNC2014 follows a sampling frame – not in the technical sense of that term, but rather in the sense of an explicitly defined schema describing (a) the registers of text from which the corpus will be sampled, as well as (b) the proportional size of each register section, that is to say, what fraction of the corpus it will ultimately make up. The design of the sampling frame for this corpus was shaped by the two competing demands made by prospective users: first, that it should be representative of the British English of the 2010s; and second, that it should maintain comparability with the original BNC, hereafter to be known as the BNC1994, to facilitate diachronic studies. These desiderata necessarily conflict: an optimally representative corpus is likely not to be optimally comparable with the earlier data.

The Spoken BNC2014 (Love, Dembry, Hardie, Brezina, & McEnery 2017), which consists of ten million running words (tokens), represents informal speech gathered across the UK (but mainly England) (Love forthcoming). This representation flows from a key decision made early in the creation of the Spoken BNC2014, namely to collect only data which occurred in informal contexts – that is, to build a corpus comparable only to the Spoken BNC1994's demographically-sampled component (hereinafter Spoken BNC1994DS) and not to the somewhat larger context-governed component or to the combination of both. This decision will be treated at greater length in Section 3. The Written BNC2014 (Hawtin forthcoming), with approximately 100 million words, forms the bulk of the overall BNC2014 (as was, of course, also the case for the BNC1994) and represents a wide range of written registers of current British usage.

Returning to the dilemma of representativeness vs. comparability, which corpus compilers are often faced with, we propose (and demonstrate later in this article) that register, as a crucial aspect of the sampling frame of the BNC2014, can help us resolve this tension. Registers, both traditional and emerging, can be considered as tools for managing and categorising a large amount of functional variation, and provide points of comparison. General corpora such as the BNC2014 are thus often structured around register-based subcorpora, which the analyst needs to consider when interpreting the results of comparative analyses.

We draw upon Biber and Conrad's (2009: 40) framework for analysing situational characteristics of registers to show that several useful situational characteristics of the Spoken BNC2014 texts have been captured by the text-level metadata and can be used for meaningful analysis at the level of register, such as multidimensional analysis (Biber 1988, 2004) and text type analysis (Biber 1989). The focus of the article is on the discussion of the functional variation in the Spoken BNC2014 with the aim of providing the necessary context for analysing register in the corpus.

3. The Spoken BNC2014: Design

The spoken component of the BNC1994 (e.g., Crowdy 1995) has for some time been "one of the biggest available corpora of spoken British English" (Nesselhauf & Römer 2007: 297). It was designed in two parts: the demographically-sampled part (c. 40%) and the context-governed part (c. 60%) (Aston & Burnard 1998). This design was, of course, the central and primary influence on all decisions that we made in creating the Spoken BNC2014. As mentioned in Section 2, we decided to focus on collection of informal speech, resulting in a dataset directly comparable in terms of register to the Spoken BNC1994DS. We made no attempt to replicate the context-governed part (spoken text recorded at meetings, appointments, educational events, and other contexts with defined goals or participant roles).

While this meant ceding grounds in terms of overall comparability, there were three very good reasons for this approach. Firstly, the sole interest of Cambridge University Press was informal speech; this was the register of British English in the Cambridge English Corpus that most urgently needed updating at the time of the project's launch. Secondly, taking into account the relatively modest budget with which we were to construct the corpus, it was most convenient to create only one 'production line' for the collection, transcription and processing of data. To have collected data from other contexts would have necessitated resources beyond our budget and would have rendered the planned size of the corpus prohibitively expensive. Thirdly, a consideration of the changed research

context in corpus linguistics and language studies since the 1990s made clear that focusing on informal conversation was the most appropriate use of our resources. Today, researchers who wish to study spoken British English occurring in specific contexts, especially public contexts, may collect their own, specialized corpora. Moreover, some such specialized corpora have been released publicly by their creators and are available to researchers with an interest in the defined context in question; examples include:

- the *British Academic Spoken English Corpus (BASE)*, which contains university lectures and seminars (Thompson & Nesi 2001);
- the *Cambridge and Nottingham Business English Corpus (CANBEC)* (Handford 2007);
- the *Characterizing Individual Speakers (CHAINS) corpus*, which represents a variety of speech styles (Cummins, Grimaldi, Leonard, & Simko 2006);
- the *Nottingham Health Communication Corpus* (Adolphs, Brown, Carter, Crawford, & Sahota 2004);
- the *Vienna-Oxford International Corpus of English (VOICE)*, which comprises face to face interactions between speakers of English as a lingua franca (Seidlhofer et al. 2013); and,
- the *TV and Movie corpora*, which contain samples of fictional speech from English-language television shows and movies (Davies 2019).

Hence, researchers with an interest in context-governed English speech already have some options open to them.[1] However, for individual researchers working on a relatively small budget, a large general corpus of informal speech, in private contexts (i.e., the register of 'informal conversation'), is harder to collect due to the requirements of size and demographic spread and the practical and ethical difficulties involved in accessing the communicative context in question. It may fairly be said, then, that the need in the field for large, publicly available datasets is substantially more acute in the case of informal speech than in the case of 'context-governed' communication. We made addressing that need the primary goal for the Spoken BNC2014.

The Spoken BNC2014 consists of over ten million running words (tokens) of informal British English conversation and is available via Lancaster University's CQPweb server (Hardie 2012) as well as file download. As with the written corpus,

1. This is generally the case; however, we acknowledge that some spoken registers which were captured in the Spoken BNC1994 (e.g., doctor-patient interaction) are yet to have been gathered and made publicly-available more recently. So, it is not the case that other contemporary corpora have fully captured every 'context-governed' register in the years since the release of the Spoken BNC1994.

it is possible to use the corpus metadata to create subcorpora for, among other purposes, direct diachronic comparison with the spoken part of the BNC1994 (see Love & Anthony in preparation). This corpus metadata exists both at the level of the text (i.e., situational variables representing features of the recording as a whole) and at the sub-text level (i.e., social variables at the level of the speaker).

Starting with the sub-text-level (i.e., social) metadata, although social dialect variation is not relevant to an analysis of register variation, Biber and Conrad (2009: 41) state that "characteristics of the speaker should be considered as part of the larger situational context for a register". The following categories are used to record metadata for the social characteristics of the speakers in the spoken BNC2014:

- exact age
- age range (BNC1994 scheme)
- age range (new scheme)
- gender
- nationality
- birthplace
- birth country
- L1
- linguistic origin
- accent/dialect as reported
- city/town living
- country living
- duration living in that location
- dialect (categorized)
- highest qualification
- occupation
- Social Grade
- NS-SEC
- L2
- foreign languages spoken
- part of core set of speakers

For further explanation of these categories, please consult the BNC2014 user guide[2] and the Spoken BNC2014 citation paper (Love et al. 2017).

Frequency information for the categories *age*, *dialect*, *gender* and *Social Grade* is presented in Tables 1–4.

---

2. Available at: <http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf> (last accessed February 2019).

Table 1. Speaker, token and utterance counts for age groups in the Spoken BNC2014

Age range	No. speakers	No. tokens	No. utterances
0–10	7	144,273	22,901
11–18	42	696,919	76,293
19–29	250	4,192,327	416,726
30–39	89	1,661,114	171,927
40–49	76	1,630,520	162,078
50–59	77	1,166,898	122,401
60–69	65	1,065,119	114,658
70–79	33	575,721	70,823
80–89	19	119,823	14,628
90–99	4	84,913	10,087
Unknown	9	84,979	14,569

Table 2. Speaker, token and utterance counts for dialect groups in the Spoken BNC2014

Dialect	No. speakers	No. tokens	No. utterances
England – Midlands	53	1,025,304	91,751
England – North	181	2,208,480	226,138
England – South	226	4,982,755	539,806
Northern Ireland	1	861	79
Non-UK	11	129,109	12,554
Republic of Ireland	6	29,907	3,363
Scotland	9	33,101	3,955
Wales	17	201,257	22,973
Unspecified	167	2,811,832	296,472

Table 3. Speaker, token and utterance counts for gender groups in the Spoken BNC2014

Gender	No. speakers	No. tokens	No. utterances
Female	365	7,072,249	725,226
Male	305	4,348,982	471,572
N/A (multiple) *	1	1,375	293

* This classification is used only for groups of multiple speakers – for instance, when providing an attribution for vocalisations that are produced by several speakers at once; a typical example would be the points found between utterances in many texts at which many or all participants in the conversation laugh together at the same time.

Table 4. Speaker, token and utterance counts for socio-economic status groups in the Spoken BNC2014

Socio-economic status	No. speakers	No. tokens	No. utterances
1.1	12	267,251	23,489
2.2	89	1,672,342	168,034
2	149	2,919,177	290,375
3	57	1,340,409	124,777
4	16	169,957	17,248
5	14	176,686	19,821
6	38	547,223	62,935
7	15	87,332	9,554
8	91	1,546,711	180,159
* (uncategorised)	169	2,308,621	258,027
Unknown	22	386,897	42,672

Tables 1 to 4 demonstrate the social diversity of the speakers in the corpus. The raw speaker metadata was provided by speakers themselves via a questionnaire which was attached to the consent form that speakers were required to sign prior to recording conversations. In several cases, we abstracted further items of metadata by classifying or normalising questionnaire responses (e.g., to abstract socio-economic status from the occupations reported in the questionnaires). We aimed to gather much more substantial metadata than is available in the BNC1994; in that corpus, there are many gaps where speakers are classified as 'unknown' on one or more demographic criteria, meaning their speech is not usable in sociolinguistic research involving the parameters in question. We sought to avoid such gaps in the metadata in the Spoken BNC2014 by incorporating the metadata questionnaire into the consent form (rather than having a separate speaker log as per the BNC1994 procedure). To demonstrate the effectiveness of this approach, Table 5 compares the number of words which populate the 'unknown' groups for the demographic features of *age*, *gender* and *socio-economic status* in the Spoken BNC1994DS with the Spoken BNC2014. A considerable improvement is evidenced by the much lower percentage of tokens in 'unknown' groups in the new corpus. This substantial improvement is an indication of the success of this approach to metadata collection; the speakers are accounted for with metadata much more richly in the Spoken BNC2014. The result of our approach to social metadata collection is that users can readily access a broad range of participant information which may be relevant as part of a wider analysis of situational context in the corpus.

Table 5. Number of word tokens categorised as 'unknown' or 'info missing' for the three main demographic categories in the Spoken BNC1994DS and the Spoken BNC2014

Demographic category	Group: 'unknown'/ 'info missing'	Spoken BNC1994DS	Spoken BNC2014
Age	Token count	698,045	84,978
	% of corpus	13.92	0.74
Gender	Token count	624,857	0
	% of corpus	12.46	0.00
Socio-economic status	Token count	1,910,794	386,896
	% of corpus	38.10	3.39

Turning to the situational (text-level) metadata, we gathered two types: categorial and non-categorial. The categorial text metadata are:

- year of recording
- recording period (year and quarter)
- number of speakers
- transcription conventions
- sample release inclusion
- transcriber

We made these categories available in the restricted query menu of CQPweb, as the individual values for each text could be easily categorized into potential subcorpus categories. The non-categorial, situational metadata we recorded are:

- recording length
- recording date
- list of speaker IDs in the recording
- location of recording
- inter-speaker relationship (e.g., family members)
- topics covered
- activity description (i.e., what the speakers were doing while recording the conversation)
- selected characterisations of conversation type

The values for these metadata groups in each text cannot be easily classified, as the contributors were given considerable freedom in terms of the situational context in which they could produce recordings. As a result, the data for these categories are quite varied, and so we recorded these verbatim as text metadata which was provided on the most part by the contributors of the recordings. For users of the corpus via CQPweb, these verbatim text metadata features are not available

to search in the tool's 'restricted query' menu, as this draws only on standardised/categorised metadata categories. However, this information is accessible to users via the subcorpus creation mechanism: using the 'scan text metadata' option in the 'Create/edit subcorpora' menu (Figure 1) allows textual searches of these features to be run to identify texts for inclusion in a subcorpus.

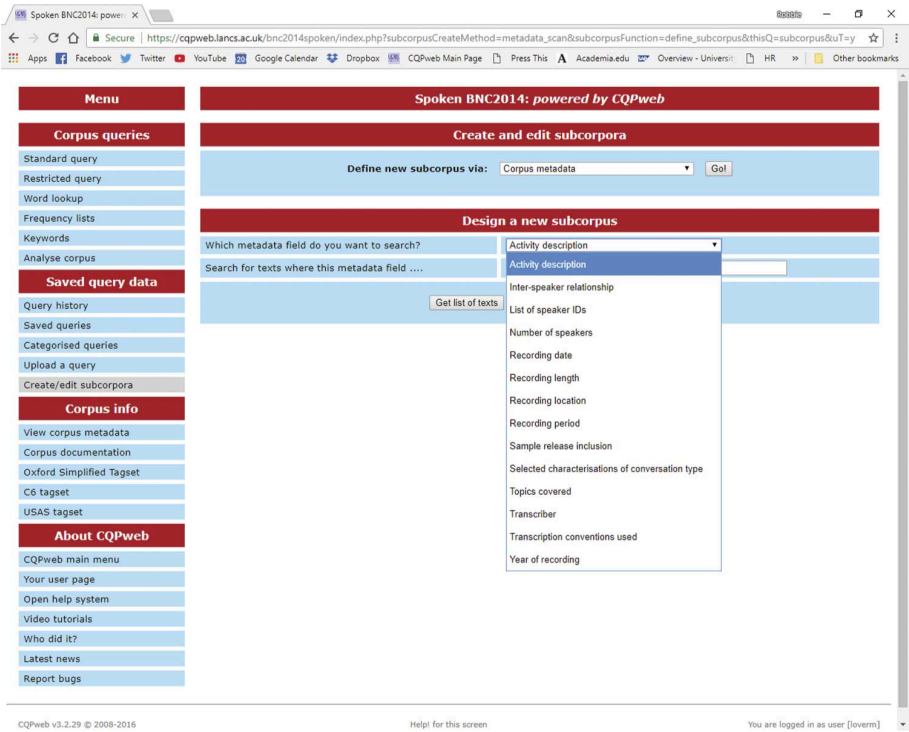

Figure 1. Creating a subcorpus based on non-categorical text metadata in CQPweb

Alternatively, all of the text-level (situational) metadata, as well as the sub-text-level (social) metadata, are available along with the XML files for users who wish to download the corpus for analysis elsewhere.

The reason for affording contributors the freedom to choose several of the situational variables was to minimize intrusiveness. The procedures noted above were already fairly intrusive, in that speakers had to go through the process of completing the informed consent form/demographic questionnaire prior to the recording. Furthermore, and unlike in the Spoken BNC1994, speakers were aware of being recorded. While it would have been possible to request contributors to make recordings in specific situational contexts so as to populate particular categories, we feared that dictating the situational context, on top of the already-

necessary imposition on speakers, would make the task too onerous (discouraging participation) or unnecessarily affect the 'naturalness' of the conversations. Thus, in terms of register, we sought to gather recordings from one broad register, i.e., informal speech. However, the result of the freedom afforded to contributors was that the corpus consists of recordings made in a large range of situational contexts, for example:

'a couple discussing modern art at a museum' (BNC2014 S23A)
'morning cuppa with mum and sister' (BNC2014 S2QU)
'grandmother's surprise visit' (BNC2014 S3CP)

It also transpired that some of the recordings had been made in locations as far from the UK as Spain, Poland, Croatia and China. No restrictions were given to contributors regarding recording location, so long as the participants were L1 speakers of British English.

The Spoken BNC2014 text metadata also includes the inter-speaker relationships, as reported by contributors. A tick-box menu was used to capture this information (i.e., *inter-speaker relationship* is categorial metadata), and so the number of texts and tokens representing each relationship can be quantified immediately (Table 6).

Table 6. Inter-speaker relationships in the Spoken BNC2014, ranked according to no. of texts

Inter-speaker relationship	Texts	Tokens
family, close friends	911	7,901,513
friends, wider family	264	2,960,614
colleagues	35	233,208
acquaintances	29	239,484
strangers	8	47,810

Most recordings were of conversations among speakers who were classified by the contributors as family or friends. This is not unexpected; this was the main target of our sampling approach, given the goal of gathering informal speech. However, over 500,000 tokens of conversation feature speakers who were colleagues, acquaintances or strangers. A useful benefit of gathering this range of text-level metadata is that several categories correspond to the characteristics in Biber and Conrad's situational framework (2009: 40).

It is clear, then, that within a corpus that captures just one register – informal speech – there is a great deal of situational variation, in terms of the physical setting of the conversation, the relationship among the speakers, and so on. It is rea-

sonable to expect that this situational variation lends itself to language variation. Therefore, within the one large register we can observe functional variation.

4. The Spoken BNC2014: Functional variation

In the rest of this paper, we put aside the sub-text metadata (the social variables pertaining to speaker demographics) and focus solely on the text metadata, i.e., the situational variables pertaining to the contexts in which recordings for the corpus were made. We consider two parameters of variation among texts – *recording location* and *activity* – both of which are categories of text metadata in the Spoken BNC2014 and help to describe the situational context of the recordings. For present purposes, we do not take into account the parameter of *inter-speaker relationship* because, as mentioned in Section 3, it has only five possible values, of which just two characterize almost all of the speakers in the corpus. By contrast, for the parameters of *recording location* and *activity* there is a much larger variety of values and, crucially, they were gathered as non-categorial metadata, i.e., free-text responses. In the search for functional variation, these metadata features seem most likely to bear fruit. The metadata fields for location and activity contained the verbatim responses of the contributors, with no attempt made prior to the release of the corpus to standardize this metadata to a finite set of category values. But, an analysis of the range of situational, and therefore functional, variation in the Spoken BNC2014 clearly required such classification of the texts. This would allow the metadata to be meaningfully aligned with Biber and Conrad's (2009) framework. Therefore, we manually classified each text according to a set of categories of situational context and activity types (Table 7), reducing each set of free-text responses to a categorical variable. The *setting* variable serves to group the reported recording locations (and corresponds directly to Biber and Conrad's 'setting' characteristic), while the *activity type* variable summarises the activity descriptions, supplemented by inferences on the basis of the recording locations where appropriate. While Biber and Conrad (2009: 40) do not explicitly include 'activity' in their framework, it is clearly relevant to several of the framework's characteristics, including 'channel', 'production circumstances' and 'communicative purposes'.

Let us consider an example of the annotation process. For corpus text S52C, the verbatim response text reads "In a bar" (for recording location) and "Friends chatting about general stuff" (for activity description). Based on this, we classified the setting for this text as 'pub/bar' and the activity type as 'general', in each case applying the closest available category from a defined, finite list. The category of 'general' is used for conversations where the speakers do not appear to be pre-

occupied by any particular activity which may structure the discourse, such as playing a game or eating a meal; texts for which even such a generalized verdict was not possible were placed in an 'unknown' category. In the case of S52C, the speakers are simply sitting in a bar and conversing with each other.

In some cases, where the activity description did not provide enough detail, inference from the recording location was the basis for the classification of the text according to activity type. So, for instance, text SV4W's activity description is "Long over-due catch-up with old friend". The activity description alone does not indicate that the speakers are having a meal during the conversation, but this is strongly implied by the fact that the response supplied for the recording location is 'Restaurant – Oxford' (which generalizes to the setting category 'pub/bar'). We therefore classified the activity type for this text as 'meal', since it unlikely that the speakers would meet in a restaurant without eating a meal. Moreover, the responses to the two queries were supplied in close proximity on a single-page form, and it is more likely that contributor would have expected these responses to be interpreted in light of one another than that they would expect each to be read independently.

The number of texts in each setting, and a summary of the activity types found in each, are reported in Table 7 (see Appendix A for individual frequencies for each setting-activity pair; e.g., 'home – general' accounts for 603 texts, while 'home – book club' accounts for only three texts).

Table 7 shows that over two-thirds of the recordings were made in participants' homes. Much as in the case of the prevalence of conversations among speakers who are related as family, this predominance is no surprise. In fact, for purposes of the overall corpus design, this was a welcome outcome; the home is the most convenient place to make a recording with family members and/or friends, and arguably the prototypical setting for the spoken register 'informal conversation'. Within the setting of the home, a range of activities have been captured, including preparing and consuming meals and playing board and console games.

That said, though, nearly a third of the recordings were *not* conducted in the home. The next most populated setting is 'vehicle'; this accounts for recordings made while travelling, mostly in cars and trains. There is also a substantial number of recordings conducted in 'office/work' settings; the majority of these appear to have been made during coffee and lunch breaks, where informal conversation, rather than task-oriented conversation is more likely to occur – and informality may characterize even 'shop talk'.

There is straightforward evidence of functional variation according to the differences in setting and activity type shown in Table 7 – which is to say, 'register' variation. Compare, for example, the family playing a board game (excerpt in

Table 7. Settings and activity types in the Spoken BNC2014 *

Setting	Activity types	No. texts	% texts
home	*book club*, *cancer support group*, *cooking*, *game*, *general*, *hair cut*, *meal*, *party*, *unknown*	872	69.65
vehicle	*general*, *unknown*	98	7.83
office/work	*general*, *meal*, *unknown*	66	5.27
pub/bar	*general*, *meal*, *unknown*	41	3.27
café	*general*, *meal*, *game*	40	3.19
outdoors	*walk*, *unknown*	37	2.96
student flat	*general*, *game*, *meal*	33	2.64
restaurant	*meal*, *unknown*	27	2.16
holiday home	*general*, *game*, *meal*	18	1.44
unknown	*unknown*, *meal*, *game*	10	0.80
hotel	*general*, *meal*	5	0.40
beauty treatment	*general*	2	0.16
church	*game*	1	0.08
museum	*general*	1	0.08
TOTAL		**1,251**	

* The full list of Spoken BNC2014 texts and their corresponding situational categorisations is available on the BNC2014 website: <http://corpora.lancs.ac.uk/bnc2014/>

Appendix B)[3] to the colleagues discussing a project during their lunch break at work (excerpt in Appendix C). In the board game text, we see, for example, shorter turns, shorter words, fewer modal verbs and a preference for first person as opposed to third person pronouns when compared to the workplace conversation. Examples of short turns (and words) in the board game text include *okay*, *mm chips*, *oh*, *it's*, *yeah*, *no* and *yeah it's*; on the other hand, the several examples of longer turns in the workplace conversation text include Examples 1 and 2:

(1) answering emails from both universities so I thought right that's it I'm not gonna answer any more

(2) right cos fir- shall we just go through some general things first.

3. The corpus extracts in Appendices B-E utilise the visualisation of the underlying XML codes for utterances, overlapping speech, etc., as is used in the current corpus setup on the Lancaster server of CQPweb.

Other differences are in the overall frequency of certain grammatical categories. The board game text contains 441 modal verbs (22,146 per million), while the workplace conversation contains 775 (27,433 per million). Similarly, the board game text contains more instances of the first-person pronoun *I* (board game: 51,122 per million vs. workplace: 36,142 per million), whereas the workplace conversation contains more instances of third person pronouns *they* (workplace: 11,363 per million vs. board game: 2,411 per million) and *them* (workplace: 5,841 per million vs. board game: 1,406 per million). These observations might prompt us, following the terminology of Biber (1988), to interpret the board game conversation as more interpersonally 'involved', and the workplace conversation as relatively more 'informational'.

Compare also the meal at the restaurant (excerpt in Appendix D) with cooking dinner at home (excerpt in Appendix E). The restaurant text features, for example, more past tense verbs and personal pronouns compared to the home cooking conversation. The restaurant text contains 230 past tense verbs (53,464 per million), while the home cooking conversation contains 58 (17,726 per million). Furthermore, the restaurant conversation contains 655 personal pronouns (152,255 per million), whereas the home cooking conversation contains 434 (132,641 per million). Since the combination of past tense verbs with personal pronouns is often drawn on for the linguistic function of narrating, we might interpret the restaurant conversation as containing more narrative than the home cooking conversation, which would be less narrative and more task-focused.

As these brief examples indicate, the availability of the situationally-defined sub-register categorizations (summarised in Table 7) make possible the analysis of the Spoken BNC2014 at the level of register. The reason for this is that each text in the corpus has now been annotated according to situational characteristics (as inspired by Biber & Conrad 2009). With sophisticated quantitative analysis, the annotations form a basis for drawing distinctions among these situationally-defined sub-registers of informal conversation according to Biber's (1988) dimensions using a multidimensional analysis approach. As shown by Biber (2004), factor and cluster analysis could be used to explore whether there is a correlation between situational characteristics (as defined by the annotation) and linguistic variation (as revealed by the analysis). A similar approach (e.g., Biber 1989) could be adopted to explore text types in the Spoken BNC2014; it would be interesting to explore whether the register of 'informal conversation' could itself be divided into text types which are distinguished according to the clustering of linguistic features across certain dimensions.

In sum, then, we have revealed the diversity of situational contexts in the Spoken BNC2014 and discussed some possible examples of functional variation in the corpus. The variety of contexts captured in the Spoken BNC2014 is a major aspect

of the value of the dataset. As well as its importance as a reference corpus of spoken English, and as a point of comparison with Spoken BNC1994DS, the Spoken BNC2014 also provides the opportunity for detailed and nuanced analysis of functional variation in contemporary informal spoken English.

5. Conclusion and further directions

In this article we have discussed how register considerations affected our approach to the compilation of the Spoken BNC2014 in the context of gathering a large representative corpus of L1 British English. For the spoken component, we sought to gather recordings from only one register ('informal conversation') but, due to differences in physical setting, activity type, and relationships among speakers, we can also use the corpus to explore functional variation across more narrowly defined register types. Our sample analyses, while merely a first step, are sufficient to demonstrate not only that such differences across texts in the Spoken BNC2014 do in fact occur, but that these differences appear to be analysable within the existing frameworks for the analysis of register variation in spoken and written language, such as the multidimensional analysis methodology pioneered by Biber (1988) and demonstrated on conversational data by Biber (2004) and text type analysis (Biber 1989).

Acknowledgements

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

References

Adolphs, S., Brown, B., Carter, R., Crawford, P., & Sahota, O. (2004). Applying corpus linguists in a health care context. *Journal of Applied Linguistics*, 1(1), 9–28.

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

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of English texts. *Linguistics* 27, 3–43.

Biber, D. (2004). Conversation text types: A multi-dimensional analysis. In G. Purnelle, C. Fairon, & A. Dister (Eds.), *Le poids des mots: Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data* (pp. 15–34). Louvain: Presses Universitaires de Louvain.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

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.

Cummins, F., Grimaldi, M., Leonard, T., & Simko, J. (2006). The CHAINS corpus: CHAracterizing INdividual Speakers. In Speech Informatics Group of SPIIRAS (Ed.), *Proceedings of SPECOM'2006 (Speech and Computer 11th International Conference)* (pp. 431–435). St Petersburg: Anatolya Publishers.

Davies, M. (2019). *The TV and Movie corpora*. Retrieved from <https://corpus.byu.edu/files/tv_movie_corpora.pdf> (February 2019).

Handford, M. (2007). The genre of the business meeting: A corpus-based study (Unpublished doctoral dissertation). University of Nottingham.

Hardie, A. (2012). CQPweb – combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17(3), 380–409.

Hawtin, A. (forthcoming). The Written British National Corpus 2014: Design, compilation and analysis (Unpublished doctoral dissertation). Lancaster University.

Love, R. (forthcoming). *Overcoming challenges in corpus construction: The Spoken British National Corpus 2014*. New York, NY: Routledge.

Love, R. & Anthony, L. (in preparation). A case for improving the textual and sub-textual analysis of corpora.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics*, 22(3), 319–344.

Seidlhofer, B., Breiteneder, A., Klimpfinger, T., Majewski, S., Osimk-Teasdale, R., Pitzl, M. -L., & Radeka, M. (2013). *The Vienna-Oxford International Corpus of English* (version 2.0 XML). <https://www.univie.ac.at/voice/page/download_voice_xml> (April 2018).

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

Appendix A. Situational contexts in the Spoken BNC2014, split into individual activity types

Situational context	Activity type	Example activity (BNC2014 text ID)	No. texts	% texts
Home	*General*	'Chatting over a cup of tea' (S2PY)	603	48.16
Home	*Meal*	'Friends eating dinner, chatting' (S7SU)	159	12.62
Vehicle	*General*	'A family conversation during a car journey' (SWZA)	78	6.23
Office/ work	*General*	'Colleagues chatting in their coffee break' (S5HH)	49	3.91
Home	*Unknown*	*No information provided*	42	3.35
Café	*General*	'Girly coffee date' (SMZV)	36	2.88

Situational context	Activity type	Example activity (BNC2014 text ID)	No. texts	% texts
Pub/bar	*General*	'Chatting in a pub' (S3UC)	36	2.88
Home	*Cooking*	'Talking whilst cooking evening meal' (SCYD)	35	2.80
Outdoors	*Walk*	'Couple go for a walk in the countryside' (S8K9)	35	2.80
Home	*Game*	'Family speaking while playing trivial pursuit' (SAZX)	28	2.24
Student flat	*General*	'Housemates talking about mobile phones' (S7DT)	28	2.24
Restaurant	*Meal*	'A meal for my father's 55th birthday' (SNJP)	26	2.08
Vehicle	*Unknown*	*No information provided*	20	1.60
Office/ work	*Meal*	'Friendly discussions over lunch break' (SFJ2)	16	1.28
Holiday home	*General*	'family discussing new house' (SHTM)	8	0.64
Holiday home	*Game*	'FAMILY PLAYING TICKET TO RIDE' (SQ6T)	5	0.40
Holiday home	*Meal*	'FAMILY EATING DINNER' (S68F)	5	0.40
Unknown	*Unknown*	*No information provided*	5	0.40
Hotel	*General*	'Two friends talking about various topics' (S4XR)	4	0.32
Pub/bar	*Meal*	'Family talking over lunch' (SKDA)	4	0.32
Unknown	*Meal*	'A couple having Sunday lunch' (SUHT)	4	0.32
Café	*Meal*	'Dinner with an old friend' (SFG3)	3	0.24
Home	*Book club*	'Talking at Book Club' (S4NB)	3	0.24
Student flat	*Game*	'Four flatmates…playing app game on iPhone' (SCWC)	3	0.24
Beauty treatment	*General*	'chatting with beautician who is a casual friend whilst having reflexology' (S28F)	2	0.16
Student flat	*Meal*	'Talking while eating pizza with housemates' (S2EF)	2	0.16
Outdoors	*Unknown*	*No information provided*	2	0.16
Café	*Game*	'Housemates talking whilst playing scrabble' (SA9P)	1	0.08
Church	*Game*	'Games for a youth group night'	1	0.08
Home	*Hair cut*	'Discussing business (badminton racket repair) while having a home haircut' (SZNP)	1	0.08
Home	*Party*	'ANON's Birthday' (SMVW)	1	0.08

Situational context	Activity type	Example activity (BNC2014 text ID)	No. texts	% texts
Home	*Cancer support group*	'Members of a cancer support group talking about their illness' (SYX3)	1	0.08
Hotel	*Meal*	'Sisters having lunch on a skiing holiday' (S3M4)	1	0.08
Museum	*General*	'A couple discussing modern art at a museum' (S24A)	1	0.08
Office/ work	*Unknown*	*No information provided*	1	0.08
Pub/bar	*Unknown*	*No information provided*	1	0.08
Restaurant	*Unknown*	*No information provided*	1	0.08
Unknown	*Game*	'Connect four competition between siblings' (SUJ8)	1	0.08

Appendix B. Excerpt of Spoken BNC2014 text S6BS, 'Family chatting while playing Trivial Pursuit' (classified in the metadata as 'Home – game')

S0602: we all start off with fifteen points
S0594: okay
S0602: if I can get these chips open
S0594: mm chips
S0602: not those kind of chips you fatty
S0594: oh
S0602: yeah that's what you get
S0594: but what did I do?
S0602: like chips is what you did
S0594: oh (.) it's a crime now
S0592: it's
S0594: >> liking chips
S0592: yeah
S0602: yeah it's
S0592: >> you're not allowed to like chips it's
S0594: no
S0592: it's been outlawed
S0602: it's a sign of capitalism
S0594: yeah it's
S0592: there's a cat there she's she's claiming the box as her own
S0600[??]: >> --UNCLEARWORD
S0602: right let's just take some scissors
S0592: your scissors
S0602: >> er (.) rude (.) see if I can
S0592: well done --ANONnameM well done

Appendix C. Excerpt of Spoken BNC2014 text SP2Y, 'Colleagues discussing a business seminar' (classified in the metadata as 'Office/work – General')

S0238: >> so did I do everything that I was supposed to do with the PowerPoint?
S0241: mm I think so let's have a look
S0240: did you send me the revised one?
S0238: er
S0241: yes but it was while ago yeah
S0238: >> well I certainly sent it to --ANONnameF
S0240: >> I'm really sorry but I did n't open emails from --ANONplace in August
S0238: >> er I would have thought I'd sent it to both of you
S0240: emails from --ANONplace in August cos I just found the first part of the holiday I was
S0238: >> no
S0240: spending my entire holiday
S0238: fair enough
S0240: answering emails from both universities so I thought right that's it I'm not gon na answer any more
S0241: no you're right
S0238: move that a little bit
S0240: that bit
S0238: and the papers
S0241: yeah confirm if you
S0240: >> so sorry I didn't yeah
S0241: >> oh we could maybe you could print it out for her or something or
S0238: well that's a
S0241: >> it's alright I mean yeah
S0238: that's a printout erm so
S0241: right cos fir- shall we just go through some general things first
S0240: >> yeah
S0241: >> that I need to confirm
S0238: >> yeah okay sorry
S0241: erm so it's confirmed that it's at the --ANONplace in --ANONplace on the twenty
S0240: >> is it a café?

Appendix D. Excerpt of Spoken BNC2014 text S29Q, 'Friends chatting about school life and friends over a snack' (classified in the metadata as 'Restaurant – Meal')

S0555: >> when my my sister me we went to the me me and my sister once when we erm we went to Sainsbury's together and we were like she was like I want to get this board of cheese and I was like is that a good idea? and I and I was like okay
S0554: >> how old were you?
S0555: no it was it was probably like last year or something

S0554: oh right
S0555: and she got it and I tried all of the cheese and it all taste horrible it tastes all of it tastes as bad as it smells
S0405: yeah
S0554: I love the flavour I like it so much
S0405: >> do you have cheese in like Malaysia?
S0555: do we have cheese in Malaysia?
S0405: how much do you consume cheese in Malaysia?
S0554: >> we have a lot of cheese here that's like cheddar cheese
S0555: >> well my m- well my mum is lactose intolerant
S0554: oh
S0405: >> oh okay
S0555: no but I'm not lactose intolerant
S0554: no no but that but that like I'm not vegetarian but I still have an intolerance to most meats
S0555: do you?
S0554: yeah
S0405: do you?
S0554: when erm when I went to the pub with my friends with my family I had like some fish I'd never had and I got so ill from it cos my body has never had it even though I eat meat my body's never had it
S0405: no give me one example fish could be bad
S0555: but the thing is fish fish fish can easily be overcooked
S0554: >> most fish and prawn prawn every time I have prawn I get sick like good prawn not bad prawn I will be sick without fail

Appendix E. Excerpt of Spoken BNC2014 text SHP3, 'Making dinner, showing ANON how to cook rice.' (classified in the metadata as 'Home – Cooking')

S0417: pour that in pour that in the saucepan
S0416: okay so pour about a third of rice and
S0417: yeah that's probably too much to be honest and then go and fill up
S0416: it might be
S0417: fill up the water
S0416: with the kettle?
S0417: no fill up fill up that with cold water
S0416: like to the brim?
S0417: >> to be to be the same amount of the as the rice
S0416: do you mean like a third?
S0417: yeah whatever the rice went up to put that up to
S0416: --UNCLEARWORD think that's --UNCLEARWORD?
S0417: --UNCLEARWORD bit more I think
S0416: oh

S0417: okay er that's as much water you need --UNCLEARWORD like as much water as there is rice
S0416: --UNCLEARWORD
S0417: put it on the thing this ignites it and then you put the lid on and you've got ta wait until it it's fifteen minutes from now
S0416: so I will go online and I'll time it
S0417: yeah
S0416: so so how do you how do you turn on this?
S0417: >> so it's three thirty-five but you have to watch it because it boils over really quickly
S0416: how do you turn on the thing? cos all you just did is click the --UNCLEARWORD
S0417: you put put it in and turn it that puts the gas on
S0416: yeah
S0417: and you ignite it
S0416: and then that's that?
S0417: yeah so you've got ta like not just leave it without putting the ignition on otherwise the gas is erm gas is dangerous without being ignited
S0416: >> so you've got ta got ta be like this and then you go like this this and then click that
S0417: it er it's the other way I think
S0416: >> or whatever posi- whatever position it's in
S0417: mm it's this
S0416: --UNCLEARWORD

Address for correspondence

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

Co-author information

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

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

Abi Hawtin
ESRC Centre for Corpus Approaches to Social Science (CASS)
Faculty of Arts and Social Sciences
Lancaster University
a.hawtin@lancaster.ac.uk

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

Claire Dembry
Cambridge University Press, University Printing House
Cambridge University Press
cdembry@cambridge.org