

The British National Corpus 2014:

User manual and reference guide

(version 1.1)



November 2018



Contents

| | |
|--|-----------|
| Contents | i |
| 1 Introduction: what is the BNC2014? | 1 |
| 2 The BNC2014 team | 2 |
| 3 Accessing the corpus..... | 2 |
| 4 The Spoken BNC2014 | 3 |
| 4.1 Data collection: recruitment & recording..... | 3 |
| 4.2 Metadata in the Spoken BNC2014 | 4 |
| 4.2.1 Metadata collection: procedure and ethics..... | 4 |
| 4.2.2 Text metadata: categories | 9 |
| 4.2.3 Text metadata: other | 14 |
| 4.2.4 Sub-text metadata: categories..... | 15 |
| 4.2.5 Speaker metadata: categories | 18 |
| 4.2.6 Speaker metadata: other | 31 |
| 4.3 Transcription | 34 |
| 4.3.1 General approach | 34 |
| 4.3.2 Main features of the transcription scheme | 35 |
| 4.3.3 Transcription procedures and quality control | 41 |
| 4.3.4 Speaker identification..... | 42 |
| 5 The Written BNC2014 | 46 |
| 6 The BNC2014 on CQPweb | 47 |
| 6.1 The Spoken BNC2014 in CQPweb..... | 47 |
| 6.2 The Written BNC2014 in CQPweb | 56 |
| 7 Encoding and markup | 57 |
| 7.1 Overall XML structure..... | 57 |
| 7.2 Spoken corpus document XML..... | 57 |

| | | |
|----------|---|-----------|
| 7.3 | Spoken corpus header XML..... | 62 |
| 7.4 | Written corpus document XML..... | 63 |
| 7.5 | Written corpus header XML..... | 63 |
| 7.6 | Changes made to the XML for annotation and CQPweb indexing..... | 63 |
| 8 | Annotation | 65 |
| 8.1 | POS tagging..... | 65 |
| 8.2 | Lematisation | 67 |
| 8.3 | Semantic tagging..... | 67 |
| 8.4 | XML for annotation | 68 |
| | References | 70 |
| | List of appendices | 73 |
| | Appendices | 74 |

I Introduction: what is the BNC2014?

The ESRC-funded Centre for Corpus Approaches to Social Science (CASS)¹ at Lancaster University is leading the compilation of the British National Corpus 2014. This is the first publicly-accessible corpus of its kind since the original British National Corpus,² which was completed in 1994, and which, despite its age, is still used as a proxy for present-day English in research today. Like its predecessor, the new corpus contains examples of written and spoken British English, gathered from a range of sources. To gather the spoken component, CASS worked together with the English Language Teaching group at Cambridge University Press (CUP), compiling a new, publicly-accessible corpus of present-day spoken British English, gathered in informal contexts. This spoken component is known as the Spoken British National Corpus 2014 (Spoken BNC2014; Love et al. 2017). The new spoken corpus contains data gathered in the years 2012 to 2016. As of September 2017 it is available publicly via Lancaster University's CQPweb server (see Hardie 2012); the underlying XML files have been downloadable from Autumn 2018 onwards. The Spoken BNC2014 contains 11,422,617 words³ of transcribed content, featuring 668 speakers in 1,251 recordings. The Written BNC2014 is currently under development (see <http://cass.lancs.ac.uk/bnc2014>).

In this guide, we use the following naming conventions for the old and new British National Corpora:

- Original BNC = 'the BNC1994'
- New BNC = 'the BNC2014'
- Spoken components = 'the Spoken BNC1994' and 'the Spoken BNC2014'
- Demographically-sampled component of the Spoken BNC1994 = 'the Spoken BNC1994DS'
- Written components = 'the Written BNC1994' and 'the Written BNC2014'

¹ The research presented in this manual was supported by the ESRC Centre for Corpus Approaches to Social Science, ESRC grant reference ES/K002155/1. Additional information on CASS and its research can be found at <http://cass.lancs.ac.uk/>

² <http://www.natcorp.ox.ac.uk/>

³ All corpus or subcorpus sizes in this document are given as counts of tokens, using the tokenisation output from the CLAWS tagger. These are the figures that can be accessed in the corpus as available via the CQPweb interface. There are two particular points of note regarding these word counts: (a) punctuation marks are counted as tokens; (b) clitics that CLAWS separates out from the bases to which they are orthographically joined, e.g., *n't*, *'m*, *'re*, *'d*, *'ve*, are also counted as tokens. Other calculation methods may produce somewhat different token counts (usually lower).

2 The BNC2014 team

Robbie Love (Lancaster University) was lead researcher for the Spoken BNC2014. Abi Hawtin (Lancaster University) is lead researcher for the Written BNC2014. The team also includes (at Lancaster University) Tony McEnery, Vaclav Brezina, Andrew Hardie, Elena Semino and Matt Timperley; and (at Cambridge University Press) Claire Dembry.

In her work on the Spoken BNC2014, Claire Dembry was supported by her team at CUP, including Olivia Goodman, Imogen Dickens, Sarah Grieves and Laura Grimes, who did much of the front-line work on the project.

An extensive team at Lancaster University and elsewhere have contributed, and continue to contribute, to the Written BNC2014. *A full set of credits will be included in the future version of this manual which accompanies the Written corpus's release.*

The construction of the Spoken BNC2014 was jointly funded by CASS and CUP. The construction of the Written BNC2014 is funded by CASS.

This corpus manual has been compiled from a combination of material published elsewhere in outputs describing the spoken and written corpora, and the project team's hitherto unpublished internal technical documentation. As such, we do not consider this manual to be a citable document separate from the BNC2014 as a resource; if you wish to refer to the contents of the manual, please cite the corpus as a whole – using one or both of the canonical references specified in the respective end-user licences for the Spoken and Written components. For example: “cf. Spoken BNC2014 (Love et al. 2017), corpus manual section 2)”.

3 Accessing the corpus

The BNC2014 is a publicly-accessible language resource, but it is not in the public domain. It remains under copyright, and use of it is subject to the terms of the User Licence. Users must agree to the licence in order to access the corpus online or to download a copy of the data. Licensing and distribution is managed by an online system accessible at <http://corpora.lancs.ac.uk/bnc2014>.

The Written and Spoken components have different copyright statuses, and therefore are subject to different licences. See Appendix A for the Spoken BNC2014 user licence. The Written BNC2014 user licence will be added when the corpus is released.

4 The Spoken BNC2014

4.1 Data collection: recruitment & 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 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 CUP team via email with further instructions. The primary method of capturing public attention was a series of national media campaigns in 2014 and 2015. Using an initial two-million-word collection made collected by CUP in 2012, we produced lists of words which had increased (e.g. 'awesome') and decreased (e.g. 'marvellous') in frequency to the greatest extent in the new data relative to the Spoken BNC1994DS. These lists were used as the basis for research press releases, which proved very popular in the national UK press. The consequent media coverage generated the most substantial intake of new contributors.

In addition to these national media campaigns, we also participated in public engagement events such as the Cambridge University Festival of Ideas (Dembry & Love 2014) and the UK Economic and Social Research Council's Festival of Social Sciences (Love 2015), where we shared early findings from a subset of the corpus and encouraged audiences to participate. Some supplementary targeted recruitment was conducted when the research team identified 'holes' in the data. Methods included use of targeted social media advertisements (e.g. targeting Facebook users from Cardiff), press releases specific to a particular social group (e.g. "Mum's the word...both then and now") and contacting colleagues from universities in sought-after locations to them to spread word of the project.

While the CUP team initiated and maintained direct contact with the contributors (i.e those who recorded conversations), they did not make any direct contact with other speakers included in the recordings. Instead, speakers received information about the project from the contributors. Contributors were therefore responsible for:

- obtaining informed consent and collecting demographic metadata from the speakers; and,
- submitting data and recordings to CUP at the end of the collection period.

Because of the importance of the contributors to the success of the project, we incentivized participation by offering 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. To ensure that all information and consent was captured, no payments were made to contributors until all metadata, consent forms and related documentation was fully completed for each recording.

Contributors were instructed to make recordings using their smartphones. They were instructed to make recordings in MP3 format (the standard format for most smartphone recording devices), and encouraged to make their recordings in fairly quiet locations, for example household interactions or conversations in quiet cafes. However, contributors were not ‘disallowed’ from recording at any time or place, since we did not want to anticipate the production of bad recordings, and advise contributors against making them, before finding out whether they would be useable. Contributors were given no restriction on the number of speakers that could be involved in conversations, although a recommendation of two to four speakers was given. Likewise, we did not impinge more than necessary upon the spontaneity of the recording sessions by dictating features such as conversation topic, although a list of suggestions was provided (see Appendix B). Finally, it was stressed to contributors that under no circumstances could they make recordings surreptitiously, and that all speakers in the conversation must be aware that recording was taking place beforehand.

4.2 Metadata in the Spoken BNC2014

4.2.1 Metadata collection: procedure and ethics

The collection of metadata is an extremely important step in the compilation of a spoken corpus as it affords the definition of subcorpora according to different features of the speakers (e.g. age) or of the recordings themselves (e.g. number of speakers in the conversation). We henceforth refer to the former type as ‘speaker metadata’ and the latter as ‘text metadata’.

Contributors were provided with copies of the Speaker Information Sheet (Figure 1), and were instructed to have each speaker fill out a copy and return it to the contributor. Since speakers had to individually sign a consent form in any case, the speaker metadata

form was incorporated into this consent form. This consent form was drafted by the team at CUP with the collaboration of the CUP legal division.



Speaker Information Sheet

(Please complete this form by hand)

1. Today's date (please write the month in words):
2. Full Name: (surname) (first)
3. Age:
4. Gender:
5. Nationality: Birthplace:
6. Mother tongue:
7. Which country has most influenced your language or the way you speak?
8. Accent / dialect:
9. Where do you currently live? (Country) (Town)
10. How many years/months have you lived there?
11. Do you speak any languages other than your mother tongue and English? ☐ YES ☐ NO
If YES, please give further details:
12. Education ☐ Secondary school ☐ College/6th Form ☐ Graduate ☐ Postgraduate
13. Job role or position:

PLEASE TURN OVER

Speaker Consent Form: British National Corpus 2014

Cambridge University Press, a department of the University of Cambridge, and Lancaster University are currently running a project to provide a resource for linguistic research. We are collecting samples of spoken British English which will be used to inform all kinds of research into the English language, and the development of teaching materials aimed at language learners. The recordings will be transcribed, anonymised, and then made into a publicly-available resource – the British National Corpus 2014.

Declaration:

1. I hereby grant to the Chancellor, Masters, and Scholars of the University of Cambridge, acting through its department Cambridge University ('Cambridge') permission to collect, store, use and otherwise exploit my/my child's writing and/or speech ('Data') and assign to Cambridge full copyright throughout the world in any resultant recordings, transcriptions and written texts.
2. I understand that the Data, and said recordings/transcriptions or extracts therefrom, may be used or licensed by Cambridge University Press for further research and development purposes and used in publications in recorded, re-recorded or written form, and I give my full consent to this use. I further understand that all citations (spoken and written) from the Data used in such publications shall be anonymised so that all references to people, places and institutions are unidentifiable.
3. I understand that any of my/my child's personal information provided as part of the Data will be stored and analysed in compliance with the UK Data Protection Act 1998. I further understand that my/my child's personal information will be used to validate and process the Data but that Cambridge will not share my/my child's personal information with any other party or use my/my child's personal information to contact me/my child for any marketing purposes, except where Cambridge may share some of my/my child's anonymised personal data, such as my/my child's age and first language, with third parties for research purposes, and sound recordings of my/my child's voice.
4. I represent and warrant that I have the full power and authority to enter into this release [on behalf of my child]; that the Data I am submitting [on behalf of my child] is original to me/my child, contains nothing libellous or unlawful and contains nothing that is in any way an infringement of any existing copyright or licence, or duty of confidentiality; the grant and other provisions of this release are not in conflict with and do not infringe any commitment, agreement or understanding that I now have or will in the future have with any other person or entity; that Cambridge's exercise of its rights under this release will not infringe the rights of any person or entity and will not cause Cambridge to incur any liability to any person or entity.
5. This release shall be interpreted in all respects in accordance with the laws of England and Wales and each party irrevocably agrees that the courts of England and Wales shall have exclusive jurisdiction to settle any dispute or claim arising out of or in connection with this release.

I further declare that:

- I am 18 years of age or older;
- All information I provide will be full and correct; and
- I give this consent freely

Name or name of minor (in block capitals)

Parent/guardian's name (in case of a legal minor):

Signature or signature of legal guardian

Date (please write name of month) e.g. 24th March 2015

Contact details (postal or email address)

Figure 1. The speaker information sheet/consent form used for collection of the Spoken BNC2014.

The gathering of metadata directly from speakers appears to have achieved its intended goal. Comparing the number of words which populate the ‘unknown’ groups of the main demographic categories in the Spoken BNC1994DS with the Spoken BNC2014 (Table 1), there has been a considerable improvement.

Table 1. Number of words categorised as ‘unknown’ or ‘info missing’ for the three main demographic categories in the Spoken BNC1994DS and the Spoken BNC2014.

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

In line with the guarantees given in the consent form, it was necessary to anonymise the data, but to accomplish this (so far as possible) in such a way as not to affect the findings of subsequent corpus analyses. These modifications included removing “references to people or places” (Baker 2010: 49), and are described in Section 4.3.2.

The second form provided to contributors was the ‘Recording Information Sheet’ (Figure 2). This information generated text metadata for the corpus. The form also includes a table in which contributors were asked to write the first turn that each speaker spoke in the corresponding recording. The purpose of this was to aid transcription; it allowed transcribers to find an example of each speaker’s voice in the recording as identified by someone who was present for the recording and likely to be familiar with each of the speakers’ voices. We collected much more text metadata than the Spoken BNC1994 team did; the speaker and text metadata categories are summarized in the next section along with their word counts in the corpus.

Recording Information Sheet

*You, the freelancer, should complete a copy of this form for each recording you make.
Please complete this form electronically*

1. Your name:
2. Date of recording (dd/mm/yy):
3. File name (e.g. BNCJLS001): BNC
4. Length of recording (hh:mm:ss):

5. Speakers on tape in order of appearance – please give the name given on their consent form, and the first words that they say, as shown in the example below:

EXAMPLE:

Speaker 1: Dave Smith, "So did you go out on Saturday....."

Speaker 1:

Speaker 2:

Speaker 3:

Speaker 4:

Speaker 5:

(Please continue on a separate, correspondingly numbered sheet if there are more than 5 speakers.)

6. Where was the recording made? (please give the location, as well as village/town/city, e.g. a coffee shop, London; The Red Lion Pub, Bristol; Speaker 2's home, Manchester):

7. How well do the speakers know each other? (select one option):

| | |
|--|--|
| Close family, partners, very close friends | |
| Friends, wider family circle | |
| Colleagues | |
| Acquaintances | |

| | |
|-----------------------------------|--|
| Strangers | |
| Teacher/pupil or lecturer/student | |

(NB, Only choose the 'teacher/pupil' option if it is the only relationship which exists between the speakers. E.g. if they also happen to be friends, tick 'Friends')

Please choose whichever category seems sensible. What we want to know is the main nature of the relationships of the speakers in this conversation. E.g. if there are 4 close family members and a visitor who is an acquaintance, choose 'Close family'. If you work as colleagues but feel your relationship is more that of friends choose 'Friends'.

8. If the speakers do not fall easily into the relationship categories above, please specify the speakers and their relationships:

9. What are the topics covered in the conversation? (List all that are covered, e.g. sport, work, the internet etc).

10. Please give your recording a short title: (E.g. friends talking about TV, friendships and birthdays; talking whilst cooking a meal with housemates; having coffee with friends talking about relationships).

11. Tick any of the following that take place in this conversation:

| | |
|---------------------|--|
| Discussing | |
| Explaining | |
| Inquiring | |
| Complaining | |
| Advising | |
| Requesting | |
| Inviting | |
| Announcing | |
| Anecdote telling | |
| Making arrangements | |
| Apologizing | |
| Buying/selling | |
| Telling jokes | |

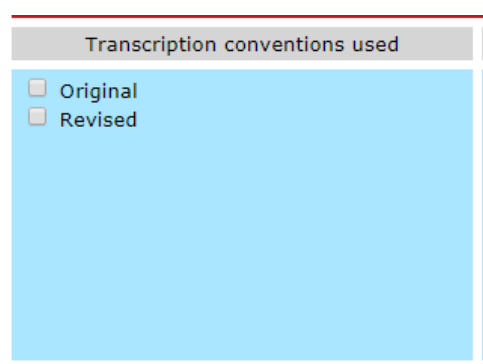
Figure 2. Recording Information Sheet used in the Spoken BNC2014.

4.2.2 Text metadata: categories

This section lists all the metadata features recorded at the level of the text that can be used to *classify* the texts into categories. For each feature, we include a brief explanation and a screenshot of the corresponding control in the CQPweb *Restricted Query* interface.

TRANSCRIPTION CONVENTIONS USED

The data in the corpus from the year 2012 was gathered by CUP before the commencement of the joint project to develop the Spoken BNC2014. The recordings from this period were therefore transcribed using conventions which are different to the transcription scheme that the research team agreed for the Spoken BNC2014 (and which is described in later sections of this manual). This initial tranche of transcriptions was automatically converted into the Spoken BNC2014 XML format. While we have made every effort to ensure that the texts derived from the 2012 recordings are formatted in the same way as the rest of the corpus, we accept that there remains a possibility of minor inconsistencies of transcription practice and/or of the use of transcription conventions. Therefore, we have made it possible to restrict queries according to which version of the transcription conventions was used to create each text.



| Conventions | No. texts | No. words |
|-------------|-----------|-----------|
| original | 220 | 2,068,054 |
| revised | 1,031 | 9,354,563 |

SAMPLE RELEASE INCLUSION

In 2016, we released a 4,789,185 word sample of Spoken BNC2014 data to a small number of researchers selected via an open application process. This sample, known as the Spoken BNC2014S (where S abbreviates *Sample*), contained all texts from the first stage of data

collection which had already been transcribed and converted into XML (see McEnery et al. 2017 for more information). These researchers were given exclusive early access to this sample via Lancaster University's CQPweb server for the purpose of conducting research projects, as proposed in their applications. In order to facilitate further work on this subset, we have made it possible to restrict queries according to whether or not texts in the full corpus were included in the Spoken BNC2014S.

Sample release inclusion

☐ Texts not in Sample release
☐ Texts within Sample release

| Sample release inclusion | no. texts | no. words |
|--------------------------|-----------|-----------|
| not in sample release | 684 | 6,633,730 |
| within sample release | 567 | 4,788,887 |

NUMBER OF SPEAKERS

This was established by counting the number of speakers listed by the contributor on the Recording Information Sheet, and subsequently checked by automated counts of the numbers of different speaker ID codes found in each transcription (excluding any instances of codes indicating an unknown speaker).

Number of speakers

☐ 12
☐ 2
☐ 3
☐ 4
☐ 5
☐ 6
☐ 7
☐ 8
☐ 9

| No. of speakers | no. texts | no. words |
|-----------------|-----------|-----------|
| two | 622 | 4,881,027 |

| | | |
|--------|-----|-----------|
| three | 335 | 3,236,340 |
| four | 198 | 2,208,413 |
| five | 54 | 564,015 |
| six | 25 | 286,496 |
| seven | 11 | 109,814 |
| eight | 2 | 70,997 |
| nine | 3 | 47,987 |
| twelve | 1 | 17,528 |

RECORDING PERIOD

This is the quarter in which recordings were gathered. Quarters are defined as 3-month periods within a given year (e.g. 2015 Q1 = January, February & March 2015).

Recording period

- ☐ 2012 Q1
- ☐ 2012 Q2
- ☐ 2013 Q3
- ☐ 2014 Q1
- ☐ 2014 Q2
- ☐ 2014 Q3
- ☐ 2014 Q4
- ☐ 2015 Q1
- ☐ 2015 Q2
- ☐ 2015 Q3
- ☐ 2015 Q4
- ☐ 2016 Q1
- ☐ 2016 Q2
- ☐ 2016 Q3

| Recording period | no. texts | no. words |
|------------------|-----------|-----------|
| 2012 Q1 | 167 | 1,558,736 |
| 2012 Q2 | 53 | 509,318 |
| 2012 Q3 | 0 | 0 |
| 2012 Q4 | 0 | 0 |
| 2013 Q1 | 0 | 0 |
| 2013 Q2 | 0 | 0 |
| 2013 Q3 | 3 | 18,052 |
| 2013 Q4 | 0 | 0 |
| 2014 Q1 | 1 | 4,247 |
| 2014 Q2 | 1 | 7,092 |
| 2014 Q3 | 183 | 1,840,798 |
| 2014 Q4 | 143 | 1,392,277 |
| 2015 Q1 | 95 | 1,009,407 |
| 2015 Q2 | 76 | 683,894 |
| 2015 Q3 | 115 | 797,787 |

| | | |
|---------|-----|-----------|
| 2015 Q4 | 161 | 1,445,885 |
| 2016 Q1 | 181 | 1,520,186 |
| 2016 Q2 | 71 | 629,861 |
| 2016 Q3 | 1 | 5,077 |
| 2016 Q4 | 0 | 0 |

YEAR OF RECORDING

This is the year in which recordings were gathered. The years are exact supersets of the quarters.



| Year | no. texts | no. words |
|------|-----------|-----------|
| 2012 | 220 | 2,068,054 |
| 2013 | 3 | 18,052 |
| 2014 | 328 | 3,244,414 |
| 2015 | 447 | 3,936,973 |
| 2016 | 253 | 2,155,124 |

TRANSCRIBER

The Spoken BNC2014 was transcribed by a total of 20 transcribers at CUP. We have included in the text metadata an anonymised identification code for the transcriber who created each text. This facilitates the investigation of possible inter-transcriber inconsistency.

| Transcriber | |
|------------------------------|--|
| <input type="checkbox"/> T01 | |
| <input type="checkbox"/> T02 | |
| <input type="checkbox"/> T03 | |
| <input type="checkbox"/> T04 | |
| <input type="checkbox"/> T05 | |
| <input type="checkbox"/> T06 | |
| <input type="checkbox"/> T07 | |
| <input type="checkbox"/> T08 | |
| <input type="checkbox"/> T09 | |
| <input type="checkbox"/> T10 | |
| <input type="checkbox"/> T11 | |
| <input type="checkbox"/> T12 | |
| <input type="checkbox"/> T13 | |
| <input type="checkbox"/> T14 | |
| <input type="checkbox"/> T15 | |
| <input type="checkbox"/> T16 | |
| <input type="checkbox"/> T17 | |
| <input type="checkbox"/> T18 | |
| <input type="checkbox"/> T19 | |
| <input type="checkbox"/> T20 | |

| Transcriber | no. texts | no. words |
|-------------|-----------|-----------|
| T01 | 8 | 186,244 |
| T02 | 162 | 1,502,359 |
| T03 | 32 | 262,640 |
| T04 | 49 | 319,645 |
| T05 | 3 | 16,221 |
| T06 | 32 | 252,091 |
| T07 | 2 | 14,294 |
| T08 | 2 | 7,841 |
| T09 | 179 | 1,734,761 |
| T10 | 351 | 3,654,617 |
| T11 | 64 | 742,584 |
| T12 | 2 | 15,399 |
| T13 | 15 | 117,093 |
| T14 | 3 | 20,933 |
| T15 | 181 | 1,669,330 |
| T16 | 1 | 2,570 |
| T17 | 2 | 10,679 |
| T18 | 86 | 439,509 |
| T19 | 48 | 218,625 |
| T20 | 29 | 235,182 |

4.2.3 Text metadata: other

This section lists all the metadata features recorded at the level of the text that *cannot* be used to classify the texts. These features do not establish categories of text, because the values can be (and often are) different for every text.

ACTIVITY DESCRIPTION

This was collected from a free-text box prompting contributors to describe what was happening while the recording was made (e.g. ‘Couple take a walk and have a chat in the countryside’). Responses to this prompt are reported as verbatim from the contributors.

INTER-SPEAKER RELATIONSHIP

This was collected from a tick-box list prompting contributors to select, in general, how well the speakers know each other (e.g. ‘friends, wider family circle’). All options selected are reported in a list in this metadata field (comma-separated).

LIST OF SPEAKER IDs

This metadata feature lists the speaker IDs which appear in the corpus text. For example, text S23A features speakers S0021, S0032, S0094, and S0095. Speaker IDs on the list are separated by spaces.

RECORDING LENGTH

The length of the recording from which the corpus text was derived. This is expressed as *H:MM:SS*, e.g. text S23A is of length 1:50:43 (one hour, fifty minutes, and forty-three seconds).

RECORDING LOCATION

This was collected from a free-text box prompting contributors to report the location where the recording was made. Unlike the Spoken BNC1994, where contributors continued recording over the length of a whole day, and so gathered data in several locations per recording session, the recording procedure for the Spoken BNC2014 assumed that each recording would take place in only one location. Responses to this prompt are reported as verbatim from the contributors.

SELECTED CHARACTERISATIONS OF CONVERSATION TYPE

This was collected from a tick-box list prompting contributors to identify the conversational acts which have taken place in the recording. Options included ‘discussing’, ‘explaining’, ‘inquiring’, ‘advising’, ‘anecdote telling’ and ‘making arrangements’. All options selected are reported in a list in this metadata field (comma-separated).

TOPICS COVERED

This was collected from a free-text box prompting contributors to list each topic covered in the conversation. Text S23A, for example, contains the following topics: ‘Computer programming, food, wine, temperature, saunas, opening presents’. Responses to this prompt are reported as verbatim from the contributors.

4.2.4 Sub-text metadata: categories

The CQPweb system uses the term *Sub-text regions* to refer to regions of the corpus texts that are enclosed within particular XML elements in the underlying transcription. The XML tags themselves are described in later sections of this manual. Here, we list the metadata features that establish categories of sub-text regions. (Other features of the XML of the corpus, which do not establish categories of sub-text regions, are described in Section 7.1.)

FOREIGN WORDS: LANGUAGE

Occasionally, speakers uttered words which were not in the English language. Transcribers were trained to recognise and mark up these words, indicating the language from which the uttered word originated. It is possible to search for any words which have been marked up in this way, according to the language spoken. This also allows users to exclude all words which have been marked as foreign. The three-letter codes shown below are from ISO 639-2/B, a standardized nomenclature for unambiguous reference to languages. This corresponds to the `<foreign>` XML tag and its *lang* attribute.

| Language | |
|--------------------------|-----|
| <input type="checkbox"/> | ara |
| <input type="checkbox"/> | bul |
| <input type="checkbox"/> | cat |
| <input type="checkbox"/> | chi |
| <input type="checkbox"/> | dut |
| <input type="checkbox"/> | fre |
| <input type="checkbox"/> | ger |
| <input type="checkbox"/> | gre |
| <input type="checkbox"/> | heb |
| <input type="checkbox"/> | hrv |
| <input type="checkbox"/> | ita |
| <input type="checkbox"/> | jpn |
| <input type="checkbox"/> | kor |
| <input type="checkbox"/> | lat |
| <input type="checkbox"/> | mao |
| <input type="checkbox"/> | ota |
| <input type="checkbox"/> | pol |
| <input type="checkbox"/> | por |
| <input type="checkbox"/> | rom |
| <input type="checkbox"/> | rus |
| <input type="checkbox"/> | spa |
| <input type="checkbox"/> | srp |
| <input type="checkbox"/> | swe |
| <input type="checkbox"/> | und |
| <input type="checkbox"/> | urd |
| <input type="checkbox"/> | wel |
| <input type="checkbox"/> | zul |

| ISO code | language | no. texts | no. words |
|----------|------------|-----------|-----------|
| ara | Arabic | 7 | 37 |
| bul | Bulgarian | 1 | 3 |
| cat | Catalan | 1 | 4 |
| chi | Chinese | 11 | 50 |
| dut | Dutch | 3 | 5 |
| fre | French | 86 | 675 |
| ger | German | 29 | 156 |
| gre | Greek | 7 | 43 |
| heb | Hebrew | 1 | 10 |
| hrv | Croatian | 4 | 25 |
| ita | Italian | 18 | 162 |
| jpn | Japanese | 8 | 53 |
| kor | Korean | 9 | 58 |
| lat | Latin | 9 | 35 |
| mao | Maori | 1 | 12 |
| ota | Turkish | 6 | 59 |
| pol | Polish | 16 | 142 |
| por | Portuguese | 2 | 33 |
| rom | Romanian | 1 | 1 |
| rus | Russian | 2 | 15 |

| | | | |
|-----|----------------|----|-----|
| spa | Spanish | 64 | 622 |
| srp | Serbian | 1 | 2 |
| swe | Swedish | 4 | 16 |
| und | Undeterminable | 25 | 193 |
| urd | Urdu | 1 | 1 |
| wel | Welsh | 7 | 52 |
| zul | Zulu | 1 | 2 |

UTTERANCE: TRANSITION TYPE

Each turn was marked up according to whether or not it overlapped the turn immediately preceding it (i.e. whether or not the ‘transition’ between turns was overlapping). Transitions are labelled as either ‘overlap’ or ‘nonoverlap’. This corresponds to the *trans* attribute on the `<u>` element in the corpus XML.

Transition type

☐ nonoverlap
 ☐ overlap

| Transition type | no. utterances | no. words |
|-----------------|----------------|-----------|
| nonoverlap | 949,122 | 9,208,260 |
| overlap | 247,969 | 2,214,346 |

UTTERANCE: ATTRIBUTION CONFIDENCE

Each turn was marked up according to whether not the transcriber was confident that they had correctly identified the speaker who produced the turn. Attribution confidence is marked as ‘high’ or ‘low’. This corresponds to the *whoConfidence* attribute on the `<u>` element in the underlying XML.

Attribution Confidence

☐ high
 ☐ low

| Confidence | no. utterances | no. words |
|------------|----------------|------------|
| high | 1,167,720 | 11,251,799 |
| low | 29,371 | 170,807 |

4.2.5 Speaker metadata: categories

This section lists all the metadata features recorded at the level of the speaker that can be used to *classify* the speakers into categories. In the underlying corpus XML, the `<u>` element that represents each utterance has a *who* attribute which contains the ID code of the speaker of that turn. Utterances within the corpus can be selected according to various metadata features of the speaker that is thus identified. For each such feature, we include an explanation and a screenshot of the corresponding control in the *CQPweb Restricted Query* interface. The available metadata features are as follows.

AGE

To classify speakers by age, it was necessary to define age brackets. In the metadata of the earlier Spoken BNC1994, speakers are categorized into the following brackets:

0-14
15-24
25-34
35-44
45-59
60+
Unknown

For most of the speakers in the Spoken BNC2014 (representing collectively 10,129,079 words of the corpus) we know the exact age, as speakers were prompted to enter their actual age (e.g. '27') in a free-text box in the metadata/consent form. This information is, indeed, available as a non-category metadata feature: see following section. For these speakers, categorization according to the brackets in the BNC1994 scheme – or, indeed, any other scheme – was possible. However, 133 speakers in the Spoken BNC2014 did not provide their exact age; they were part of the initial phase of data collection in 2012, during which information about speaker age was recorded according to the following brackets, rather than the exact age:

0-10
11-18
19-29
30-39
40-49
50-59
60-69
70-79
80-89
90-99
Unknown

It was only after the 2012 phase of collection, that we decided to start collecting the exact age of speakers. Reclassification of the first-phase data according to the BNC1994 scheme was, therefore, not possible across the board. Nonetheless, we endeavoured to classify as many of these speakers as possible according to the older scheme for the sake of comparability. Those in the 50-59 category could be assigned to Spoken BNC1994DS 45-59 category, and those in the 70-79, 80-89, and 90-99 categories could be assigned to the Spoken BNC1994DS 60+ category. This accounts for 26 speakers.⁴ The remaining 107 cannot be recategorized into the old groupings, because of overlaps between the categories. The modern 19-29 category, for example, straddles the boundary between the older 15-24 and 25-34 categories, and so 19-29 speakers without exact age records cannot be placed in either with certainty. One workaround, proposed by Laws et al. (2017), is to place half of the sum of tokens from each of the straddling categories (11-18 / 19-29 / 30-39 / 40-49) into the relevant categories from the older scheme. So, the frequency of instances of a given query as produced by, for example, the 30-39 group in the Spoken BNC2014 would be divided equally between the 25-34 and 35-44 groups for comparison with the 1990s data. This maximises the amount of data in each age band, at the cost of blurring the accuracy of the distinctions between adjacent age-bands to random (and thus unknown) degree. While Laws et al.'s solution may be highly appropriate for some particular research purposes, it

⁴ Likewise, speakers in the range 0-10 without exact age could have been added to the 0-14 category; however, as it happens, there were not any speakers aged 0-10 in the 2012 phase of data collection.

could not be adopted at the level of the canonical speaker metadata without misrepresenting the nature of the corpus

Thus, the 107 speakers whose age bracket in the Spoken BNC1994DS system cannot be determined have instead been classified into the ‘Unknown’ category. This unfortunately means that over one million words of Spoken BNC2014 data are excluded from age comparisons with the Spoken BNC1994.

Overall, then, the speaker metadata includes two different features classifying the speakers by age. The first uses the brackets of the older scheme from the BNC1994, and is labelled *Age (BNC1994 groups)* in the CQPweb interface. The second uses the brackets that were employed in the initial phase of data collection in 2012, as described above, and is labelled *Age range* in the CQPweb interface. The latter scheme is designed to facilitate more fine-grained apparent-time analysis of the new data; it 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).

| Age (BNC1994 groups) | Age range |
|----------------------------------|----------------------------------|
| <input type="checkbox"/> 0-14 | <input type="checkbox"/> 0-10 |
| <input type="checkbox"/> 15-24 | <input type="checkbox"/> 11-18 |
| <input type="checkbox"/> 25-34 | <input type="checkbox"/> 19-29 |
| <input type="checkbox"/> 35-44 | <input type="checkbox"/> 30-39 |
| <input type="checkbox"/> 45-59 | <input type="checkbox"/> 40-49 |
| <input type="checkbox"/> 60+ | <input type="checkbox"/> 50-59 |
| <input type="checkbox"/> Unknown | <input type="checkbox"/> 60-69 |
| | <input type="checkbox"/> 70-79 |
| | <input type="checkbox"/> 80-89 |
| | <input type="checkbox"/> 90-99 |
| | <input type="checkbox"/> Unknown |

| Age (BNC1994 groups) | No. speakers | No. words |
|----------------------|--------------|-----------|
| 0-14 | 15 | 309,177 |
| 15-24 | 159 | 2,777,761 |
| 25-34 | 92 | 1,622,317 |
| 35-44 | 50 | 1,379,783 |
| 45-59 | 117 | 2,194,465 |
| 60+ | 121 | 1,845,576 |
| Unknown | 117 | 1,293,527 |

| Age range | no. speakers | no. words |
|-----------|--------------|-----------|
| 0-10 | 7 | 144,273 |
| 11-18 | 42 | 696,919 |
| 19-29 | 250 | 4,192,327 |

| | | |
|---------|----|-----------|
| 30-39 | 89 | 1,661,114 |
| 40-49 | 76 | 1,630,520 |
| 50-59 | 77 | 1,166,898 |
| 60-69 | 65 | 1,065,119 |
| 70-79 | 33 | 575,721 |
| 80-89 | 19 | 119,823 |
| 90-99 | 4 | 84,913 |
| Unknown | 9 | 84,979 |

DIALECT

A free-text box on the speaker metadata form prompted the speakers to report their own accent/dialect. Responses were categorized according to the *Nomenclature of Territorial Units for Statistics* (NUTS) statistical regions of the UK, created in 1994 by the John Major government. In this system, Scotland, Wales, and Northern Ireland are each treated as single regions, and England is divided into several regions. These divisions were, until 2011, used to define the Government Offices for the English Regions. Despite being abolished in 2011, the regions have continued to be used for statistical analysis by the Office for National Statistics in national surveys such as the Labour Force Survey and the Annual Population Survey since the year 2000 (ONS 2013; 2014). The regions are:

- (1) North East
- (2) North West
- (3) Merseyside⁵
- (4) Yorkshire & Humberside
- (5) East Midlands
- (6) West Midlands
- (7) Eastern
- (8) London
- (9) South East
- (10) South West
- (11) Wales
- (12) Scotland
- (13) Northern Ireland. (ONS 2014: 41)

⁵ Despite appearing in the list, “Merseyside is generally included in the North West region in published data” (ONS 2014: 41), meaning that only twelve categories are used in most surveys.

The advantage of the NUTS scheme is that it opens the door for possible alignment between the corpus data and contemporary UK population data collected by the Office for National Statistics. Our most precise grouping of speakers according to self-reported dialect was based on these thirteen regions, with additional categories for locales beyond the UK (Ireland and the rest of the world). However, as not all speaker self-reports allowed such precise classification, we provided three other levels of classification, according to a procedure that will now be described.

Based on speakers' free-text answers to the question of what variety of English they speak, each speaker is assigned to a category at each of the four levels in Figure 3 ("global", "country", "supraregion" and "region"). The assignments depend upon how much could be inferred from their self-reported response, with the aim of maximizing specificity (in other words, to "get as much out of" the metadata as possible). 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.

| (1) Global | (2) Country | (3) Supra-region | (4) Region |
|-------------|--------------------|------------------|-----------------------------|
| UK | English | North | North East |
| | | | Yorkshire & Humberside |
| | | | North West (not Merseyside) |
| | | | Merseyside |
| | | | Merseyside |
| | | Midlands | East Midlands |
| | | | West Midlands |
| | | South | Eastern |
| | | | South West |
| | | | South East (not London) |
| Non-UK | Scottish | Scottish | London |
| | | | Scottish |
| | | | Welsh |
| | | | Welsh |
| | | | Northern Irish |
| | | | Northern Irish |
| Non-UK | Irish ⁶ | Irish | Irish |
| | | | Non-UK ⁷ |
| Unspecified | Unspecified | Unspecified | Unspecified |

Figure 3. Dialect categories used in the Spoken BNC2014.



⁶ 'Irish' implies 'Republic of Ireland' as the category is mutually exclusive with 'Northern Irish'.

⁷ At all levels other than 'Global', 'Non-UK' is to be interpreted as 'Non-UK other than ROI' – since 'Irish' is a separate category, the Republic of Ireland being the only country, other than the UK, which can usefully be distinguished at any level given the actual distribution of speakers in the corpus.

| Dialect at Level 3 | Dialect at Level 4 |
|---|--|
| <input type="checkbox"/> midlands <input type="checkbox"/> n_ireland <input type="checkbox"/> non_uk <input type="checkbox"/> north <input type="checkbox"/> r_ireland <input type="checkbox"/> scotland <input type="checkbox"/> south <input type="checkbox"/> unspecified <input type="checkbox"/> wales | <input type="checkbox"/> e_midlands <input type="checkbox"/> eastern_engl <input type="checkbox"/> liverpool <input type="checkbox"/> london <input type="checkbox"/> n_ireland <input type="checkbox"/> non_uk <input type="checkbox"/> northeast <input type="checkbox"/> northwest <input type="checkbox"/> r_ireland <input type="checkbox"/> scotland <input type="checkbox"/> southeast <input type="checkbox"/> southwest <input type="checkbox"/> unspecified <input type="checkbox"/> w_midlands <input type="checkbox"/> wales <input type="checkbox"/> yorkshire |

| Dialect at level 1 | no. speakers | no. words |
|--------------------|--------------|-----------|
| non_uk | 17 | 159,016 |
| uk | 566 | 9,932,384 |
| unspecified | 88 | 1,331,206 |

| Dialect at level 2 | no. speakers | no. words |
|--------------------|--------------|-----------|
| england | 530 | 9,587,388 |
| n_ireland | 1 | 861 |
| non_uk | 11 | 129,109 |
| r_ireland | 6 | 29,907 |
| scotland | 9 | 33,101 |
| unspecified | 97 | 1,440,983 |
| wales | 17 | 201,257 |

| Dialect at level 3 | no. speakers | no. words |
|--------------------|--------------|-----------|
| midlands | 53 | 1,025,304 |
| n_ireland | 1 | 861 |
| non_uk | 11 | 129,109 |
| north | 181 | 2,208,480 |
| r_ireland | 6 | 29,907 |
| scotland | 9 | 33,101 |
| south | 226 | 4,982,755 |
| unspecified | 167 | 2,811,832 |
| wales | 17 | 201,257 |

| Dialect at level 4 | no. speakers | no. words |
|---------------------------|---------------------|------------------|
| e_midlands | 11 | 75,676 |
| eastern_engl | 27 | 823,235 |
| liverpool | 15 | 157,342 |
| london | 36 | 363,240 |
| n_ireland | 1 | 861 |
| non_uk | 11 | 129,109 |
| northeast | 27 | 409,078 |
| northwest | 62 | 752,288 |
| r_ireland | 6 | 29,907 |
| scotland | 9 | 33,101 |
| southeast | 32 | 501,397 |
| southwest | 20 | 261,365 |
| unspecified | 354 | 6,789,782 |
| w_midlands | 8 | 197,633 |
| wales | 17 | 201,257 |
| yorkshire | 35 | 697,335 |

HIGHEST QUALIFICATION

This was collected from tick-boxes prompting speakers to select their highest level of education.

Highest qualification

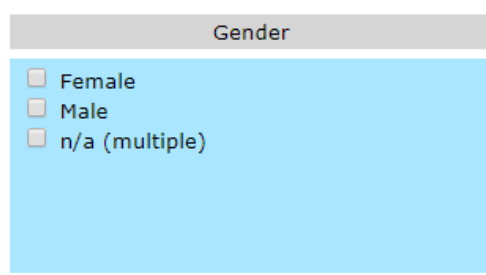
☐ Primary
☐ Secondary
☐ Sixth-form
☐ Graduate
☐ Postgrad
☐ Unknown

| Highest qualification | no. speakers | no. words |
|------------------------------|---------------------|------------------|
| 1_primary | 2 | 122,289 |
| 2_secondary | 88 | 1,279,961 |
| 3_sixth-form | 172 | 2,083,241 |
| 4_graduate | 224 | 4,405,174 |
| 5_postgraduate | 164 | 3,226,767 |
| 9_unknown | 21 | 305,174 |

GENDER

Gender was collected following a similar procedure to that reported by Crowdy (1993) for the Spoken BNC1994DS, but the 'M or F' prompt used in Crowdy's approach was omitted,

and replaced by a free-text box. In light of “the complexity and fluidity of sex and gender categories” (Bradley 2013: 22), we did not presuppose that all speakers would willingly describe their gender in this binary fashion. In fact, however, all speakers did report 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).



The image shows a web form titled "Gender" with a light blue background. It contains three radio button options: "Female", "Male", and "n/a (multiple)".

| Gender | no. speakers | no. words |
|----------------|--------------|-----------|
| F | 365 | 7,072,249 |
| M | 305 | 4,348,982 |
| n/a (multiple) | 1 | 1,375 |

PART OF CORE SET OF SPEAKERS

A set of 250 speakers was selected as a ‘core set’ of speakers. The core set is designed so that the subsection of the corpus consisting of just these speakers’ utterances has a better quantitative balance across the various social categories than does the complete BNC2014. The selection of speakers for the ‘core’ set was undertaken by Susan Reichelt as part of a project aimed at exploiting both the ‘core’ Spoken BNC2014 and a similar ‘core’ within the BNC1994DS for the purpose of variationist sociolinguistic analysis.⁸ We gratefully acknowledge this contribution to the corpus metadata.

⁸ This project, entitled ‘*The British National Corpus as a sociolinguistic dataset: Exploring individual and social variation*’, is funded by ESRC: grant reference ES/P001559/1, principal investigator Vaclav Brezina. For more information see http://cass.lancs.ac.uk/?page_id=2087.

Part of core set of speakers

☐ Not in core set
☐ In core set

| Part of core set of speakers | no. speakers | no. words |
|------------------------------|--------------|-----------|
| n | 421 | 5,253,310 |
| y | 250 | 6,169,296 |

CLASS: NS-SEC & CLASS: SOCIAL GRADE

The speaker metadata form included a free-text box prompting speakers to report their current occupation.

In the Spoken BNC1994, speakers' socio-economic status was estimated from their occupation based on the categories from the National Readership Survey's Social Grade demographic classification system (Table 2). This system has been accepted for use in the creation of UK demographic data in the market research industry for over half a century (Collis 2009: 2).

Table 2. National Readership Survey Social Grade classifications (NRS 2014).

| Code | Description |
|------|--|
| A | Higher managerial, administrative and professional |
| B | Intermediate managerial, administrative and professional |
| C1 | Supervisory, clerical and junior managerial, administrative and professional |
| C2 | Skilled manual workers |
| D | Semi-skilled and unskilled manual workers |
| E | State pensioners, casual and lowest grade workers, unemployed with state benefits only |

In 2001, an ESRC review of the existing Office for National Statistics social classifications (Rose & O'Reilly 1998) prompted the creation of a new system, the National Statistics Socio-economic Classification (NS-SEC). The nine main categories in this system of classification are given in Table 3.

Table 3. The nine major analytic classes of the *NS-SEC* (ONS 2010c).

| NS-SEC | Description |
|--------|--|
| 1 | Higher managerial, administrative and professional occupations: ⁹ |
| 1.1 | Large employers and higher managerial and administrative occupations |
| 1.2 | Higher professional occupations |
| 2 | Lower managerial, administrative and professional occupations |
| 3 | Intermediate occupations |
| 4 | Small employers and own account workers |
| 5 | Lower supervisory and technical occupations |
| 6 | Semi-routine occupations |
| 7 | Routine occupations |
| 8 | Never worked and long-term unemployed |
| * | Students/unclassifiable |

The NS-SEC is now the government standard and was used for the 2001 and 2011 censuses as well as the Labour Force Survey (LFS) (ONS 2015: 125). Compared to the Social Grade system, NS-SEC is more detailed and, therefore, may allow a more sensitive analysis of the relationship between socio-economic categories and language use; moreover, its official status means that it is compatible with a range of government datasets.

Based on speakers' reported occupation, we coded each speaker for both the National Readership Survey's Social Grade and the Office for National Statistics' NS-SEC; this facilitates the comparison of speakers between the Spoken BNC2014 and the Spoken BNC1994DS using Social Grade. Rather than separately categorise the Spoken BNC2014 speakers' occupations into Social Grade, we derived Social Grade classifications automatically from the speakers' NS-SEC codes. As there is no objective tool available for the classification of occupations according to Social Grade, in terms of consistency of judgement it seemed appropriate to automate the Social Grade classification process by mapping the NS-SEC codes onto the Social Grade codes. No formal standard has been established for translating either of these schemes to the other, but in the interests of

⁹ Category 1 is not in and of itself an analytic category; rather it comprises analytic categories 1.1 and 1.2, which can be merged to form category 1.

comparability we proposed an automatic mapping from NS-SEC to Social Grade so that both schemes can be analysed in the Spoken BNC2014 (presented in Table 4). The result is that each speaker in the Spoken BNC2014 has been assigned both an NS-SEC and Social Grade socio-economic status code.

Table 4. 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 | 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 | | |

MAPS ON TO...

| Class: NS-SEC | Class: Social grade |
|--|----------------------------------|
| <input type="checkbox"/> 1:1 | <input type="checkbox"/> A |
| <input type="checkbox"/> 1:2 | <input type="checkbox"/> B |
| <input type="checkbox"/> 2 | <input type="checkbox"/> C1 |
| <input type="checkbox"/> 3 | <input type="checkbox"/> C2 |
| <input type="checkbox"/> 4 | <input type="checkbox"/> D |
| <input type="checkbox"/> 5 | <input type="checkbox"/> E |
| <input type="checkbox"/> 6 | <input type="checkbox"/> Unknown |
| <input type="checkbox"/> 7 | |
| <input type="checkbox"/> 8 | |
| <input type="checkbox"/> * (uncategorised) | |
| <input type="checkbox"/> Unknown | |

| Class: NS-SEC | no. speakers | no. words |
|-------------------|--------------|-----------|
| 1.1 | 12 | 267,251 |
| 2.2 | 89 | 1,672,342 |
| 2 | 149 | 2,919,177 |
| 3 | 57 | 1,340,409 |
| 4 | 16 | 169,957 |
| 5 | 14 | 176,686 |
| 6 | 38 | 547,223 |
| 7 | 15 | 87,332 |
| 8 | 91 | 1,546,711 |
| * (uncategorised) | 169 | 2,308,621 |
| unknown | 22 | 386,897 |

| Class: Social Grade | no. speakers | no. words |
|---------------------|--------------|-----------|
| A | 101 | 1,939,593 |
| B | 149 | 2,919,177 |
| C1 | 73 | 1,510,366 |
| C2 | 14 | 176,686 |
| D | 53 | 634,555 |
| E | 260 | 3,855,332 |
| unknown | 21 | 386,897 |

4.2.6 Speaker metadata: other

This section lists all the metadata features recorded at the level of the speaker that *cannot* be used to classify the speakers. These features do not establish categories of speaker, because the values can be (and often are) different for every speaker. (In the CQPweb interface, these features are thus not available for use in restricted query, but are viewable from the concordance, and usable in the subcorpus creation process.)

EXACT AGE

This is the free-text response from which the different age categorisations described above were derived.

NATIONALITY

This metadata feature was collected from a free-text box prompting speakers to provide their nationality. It should be noted that this information was *not* used as a basis for including or excluding the contributions of particular speakers to this corpus of specifically *British* English; this was based instead on first language; see below.

PLACE OF BIRTH

This metadata feature was collected from a free-text box prompting speakers to provide their birthplace.

FIRST LANGUAGE

This metadata feature was collected from a free-text box prompting speakers to provide their LI. Only ten speakers in the corpus reported an LI other than (British) English. The recordings in which they feature were not excluded from the corpus because in those recordings, these speakers interacted with LI speakers of British English, and the contribution of the non-LI British English speakers was not substantial.

LINGUISTIC ORIGIN

This metadata feature was collected from a free-text box prompting speakers to report the country/countries that they believe have been most influential on their LI use.

ACCENT/DIALECT AS REPORTED

This is the free-text response from which the four levels of regional dialect categories described above were derived.

CITY/TOWN LIVING, COUNTRY LIVING, & DURATION LIVING THERE

This metadata feature was collected from free-text boxes prompting speakers to provide the town and country in which they currently live, followed by the number of years/months that they have lived there.

OCCUPATION (TITLE)

This metadata feature contains the free-text response on occupation that was used to derive socio-economic status categories as described above.

SECOND LANGUAGE & FOREIGN LANGUAGES SPOKEN

Speakers were prompted to specify what their L2 was, other than British English, if they considered themselves bilingual. Most speakers (640) gave a null response to this question. Four speakers specified Spanish. Three speakers specified each of French, German, Gujarati, and Irish. Two specified Hungarian. The following languages were specified by one speaker each: Arabic, Cantonese, Dutch, Italian, Kikuyu, Kutchi, Russian, Swedish, Turkish, Urdu, and Welsh. Two speakers specified a pair of languages: “Spanish, Swedish” and “Swedish, Welsh”.

Speakers were also prompted to list any foreign languages that they spoke, and to specify the level their level of ability. The metadata is laid out in a slightly reorganised format from the verbatim responses: the foreign language is listed, joined with a dash to the level of competence claimed (or “level unspecified” if the speaker did not give that information); where multiple languages were reported, they are given as a semi-colon delimited list. 193 speakers out of 668 gave a non-null response. Some examples of the reformatted responses are as follows:

- Spanish -- A2
- French -- School level
- French -- some; Hausa -- once fluent, now gone
- French -- Fluent; German -- A-level; Spanish -- A-level
- French -- A level
- Spanish -- Rudimentary; French -- almost conversational
- French -- Rudimentary; German -- Rudimentary; Spanish -- Rudimentary
- Welsh -- level unspecified; Turkish -- level unspecified; Arabic -- level unspecified; Zulu -- level unspecified

No attempt has been made to standardise references to languages in these two metadata fields. Instead, the languages are given verbatim as reported.

4.3 Transcription

4.3.1 General approach

The creation of the corpus involved orthographic, rather than phonetic, transcription. 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); an orthographic transcription serves the needs of research in these areas. Many differing standards exist in corpus linguistics for the transcription of audio data, including: the AHDS (Arts and Humanities Data Service) guide (Thompson 2005), CHAT (codes for the human analysis of transcripts; MacWhinney 2000), the CES (Corpus Encoding Standard; Ide 1996); the ICE (International Corpus of English) standard (Nelson 2002), the NERC (Network of European Reference Corpora)/COBUILD conventions (Payne 1995), and the Santa Barbara School conventions (Du Bois et al. 1993); see further the review by Andersen (2016) of these systems. Another system is the Text Encoding Initiative (TEI; Burnard & Bauman 2013), which was used to encode the BNC1994, although the Spoken BNC1994’s transcription scheme itself (Crowdy 1994) was comprised of simpler tags which were (eventually) converted into the TEI format. The BNC1994’s TEI structure was initially encoded using the angle-bracket-based tags of SGML (*Standard Generalized Markup Language*); later releases converted the SGML tags to XML tags (*eXtensible Markup Language*). XML is currently the preferred standard mode of corpus markup.

Following this general approach, we opted to encode the corpus following the recommendations of Hardie (2014) for a ‘modest’ level of XML markup using, by and large, similar tags to those defined by the TEI. However, we did not require transcribers to insert the XML tags at the time of transcription. Rather, we designed a more human-friendly transcription scheme based on short, easy-to-type codes for the different features to be captured, making sure that all of these short codes could be unambiguously mapped to XML at a later stage. We therefore used a set of automated conversion scripts to translate the initial transcripts into XML. 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.

The Spoken BNC2014 transcription scheme was designed taking into account certain useful recommendations made by Atkins et al. (1992: 11-12), including: beginning each turn with a code identifying 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 note that classifying functional and non-functional sounds (also known as filled pauses, or more informally *ums* and *ahs*) 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). All these recommendations have been adopted. The following section explains the transcription scheme as it was presented to the transcribers; the XML which was then generated as the canonical version of the corpus is described in Section 7.2.

4.3.2 Main features of the transcription scheme

This section discusses some of the main features of the Spoken BNC2014 transcription scheme. The examples given in this section are in the format of the corpus texts *as they were originally transcribed*, and not how they appear either in the canonical XML format of the corpus itself or in the CQPweb interface. The transcription scheme as presented here 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 made regarding the transcription process 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.

The scheme’s main features are detailed below, with examples from the transcription scheme where appropriate. The entire scheme can be found in Appendix C.

- Encoding of speaker IDs

Each speaker was given a unique numeric code (a speaker ID, e.g. ‘<0022>’), which was consistent across every recording in which they were recorded. (A leading ‘S’ was added to the speaker ID codes in the XML, to create a non-numeric unique identifier.) Transcribers were also given the facility to indicate cases where they were not fully confident in their identification of the speaker who produced a given turn, but could provide a best guess; see Section 4.3.4.

- De-identification

De-identification, or anonymization, is the process of ensuring “any reference that would allow an individual to be identified is omitted from the transcription” (Crowdy 1994: 28). We opted not to explore the use of automatic methods to de-identify speakers once the transcripts had already been produced. Such post hoc processing of names and other details is strongly recommended against by Hasund (1998: 14) in her account of anonymization procedures in COLT:

(a) Automatic replacement will wrongly affect:

- first names of public persons (actors, singers, etc.)
- inanimate objects with person names (computer games, etc.)
- nouns, adjectives, verbs, etc. that overlap in form with names (untagged corpora only);

(b) Speakers sometimes use names creatively to make puns, alliteration, rhymes, etc., the effects of which would partly or completely be destroyed by replacements;

(c) Automatic replacement is complicated by instances where the pronunciation of a name is initiated, but not completed, by a speaker (e.g. Cha= instead of Charlie).

Instead, we integrated de-identification of names into the process of transcription. Following Gablasova et al. (2015), the names of (non-famous)¹⁰ people were de-identified completely with the tag <name>, with the gender of the name added (where interpretable). The following example from the Spoken BNC2014 transcripts contains the name of a female:

<0326> I dunno what to what to do for <name F>'s birthday

Transcribers were instructed, in the case of names that are used for both males and females (e.g. “Sam” for “Samantha” or, equally, “Samuel”), to use the tag <name N>, unless the gender of the referent could be inferred from context (i.e. use of pronouns). The inclusion of gender was a crude attempt to acknowledge at least in part that “names...carry a certain amount of social and ethnic information” (Hasund 1998: 13), which could be retained without compromising anonymity.

¹⁰ Despite the attested difficulty in consistently distinguishing between the names of famous and non-famous people (Hasund 1998: 20), we encouraged transcribers to avoid anonymizing the names of celebrities, fictional characters, etc., who are considered to be in the public domain.

Aside from names, transcribers were instructed to use de-identification tags for other personally identifiable information including addresses and phone numbers (Crowdy 1994: 28; Hasund 1998: 13), and locations or institutions that seem unique to the speaker in some way. Newly emerging personal identifiers such as email address and social media usernames were also removed:

<0325> yeah someone called <soc-med> follows me it's like co= no because I was reading through basically what happens is I was then I found like a confession thing and I was just reading through all of them

- Minimal use of punctuation

Transcribers were instructed not to use most punctuation marks (full stops, commas, semi-colons, exclamation marks). It would in theory have been possible to instruct them to use such punctuation as it is used in writing, i.e. to represent (different types of) discourse boundaries (whether grammatical, pragmatic, and/or prosodic). However, we did not have any means of ensuring that transcribers would use such punctuation consistently (over time, or between transcribers). There would thus be a high risk that any punctuation would be potentially misleading to analysts. We judged it better to disallow punctuation, rather than let it be included in the corpus inconsistently. Instead of normal commas, full stops, and so on, transcribers were given two short codes to use to represent pauses specifically (and not any kind of non-pause boundary). Short pauses (up to 5 seconds) were marked as ‘(.)’ and long pauses (more than 5 seconds) were marked as ‘(...)’, as exemplified below:

<0618> yeah (.) okay because you see I thought the same about when Cameron got in again I thought holy shit I don't know anybody who's voted for this arse

<0405> Fanta is orange already (...) oh sh=

The only feature of written punctuation retained in the Spoken BNC2014 transcription scheme is the question mark, which was to mark questions (and *not* as a discourse boundary marker). This was because we had higher confidence in the transcribers' ability to accurately flag questions than in their ability to use commas, full stops and semi-colons in a consistent manner. In English, grammatically formed interrogatives can be grouped into three

categories: *yes/no questions*, *wh-questions* and *tag questions* (Börjars & Burridge 2010: 108-115). Below is an example of each question type from the Spoken BNC2014:

Yes question:

<0202> and I want to fuck you so (.) sorry did that make you feel awkward?

Wh-question:

<0202> what time is it now?

Tag question:

<0619> it's quite nice in this window isn't it?

In pilot testing, the transcribers reported that they were confident in identifying fully grammatically formed questions in the forms of these three main varieties. However, using question marks *only* for such forms appeared too restrictive; the transcribers observed that there were many more cases where they were confident that a question was being asked, but that lacked a fully grammatical interrogative form. These included questions expressed incompletely (with some surface form(s) omitted), or questions expressed in declarative form with audible rising intonation:¹¹

Incompletely formed interrogative structures:

ah is it lovely and warm there Dylan? **getting dried off?**

pardon?

mm yeah exactly sorry?

Declarative structures functioning as questions:

so he has someone there who does all this then?

how many years have we lived here? **two and a half years?**

we're talking mains?

¹¹ In examples which contain more than one question, the one which exemplifies the question type is emboldened.

These all clearly function as questions without having full interrogative forms. The transcribers reported that it was intuitively easier to include question marks in examples such as these than to exclude them and to expend time checking that only fully structurally formed interrogatives were flagged with question marks. It appears that allowing transcribers the freedom to use intuitive criteria for the coding of question marks, rather than purely structural criteria, adds useful detail to the transcription while apparently reducing transcriber effort. In consequence, this is the approach that was adopted in the transcription scheme.

- Overlaps

Where the beginning of one speaker turn overlaps in time with the end of the previous turn, it is labelled with the code . No attempt was made to indicate the beginning and end points of the overlap. Instead, the overlap is conceptualised as a feature of the transition between utterances.

- Filled pauses

To ensure the greatest possible consistency in the transcription of filled pauses (*ums* and *ahs*) we provided transcribers with a list of eight filled pause sounds, and instructed them not to use any representation other than one of these eight. The instructions for their use included phonetic guidance as well as reference information about the common discourse function of each sound (Table 5).

Table 5. List of permissible filled pauses in the Spoken BNC2014 transcription scheme.

| What it sounds like | How to write it |
|--|-----------------|
| Has the vowel found in “father” or a similar vowel; usually = realisation, frustration or pain | ah |
| Has the vowel found in “road” or a similar vowel; usually = mild surprise or upset | oh |
| Has the vowel in “bed” or the vowel in “made” or something similar, without an “R” or “M” sound at the end; usually = uncertainty, or ‘please say again?’ | eh |
| A long or short “er” or “uh” vowel, as in “bird”; there may or may not be an “R” sound at the end; usually = uncertainty | er |
| As for “er” but ends as a nasal sound | erm |
| Has a nasally “M” or “N” sound from start to end; usually = agreement | mm |
| Like an “er” but with a clear “H” sound at the start; usually = surprise | huh |
| Two shortened “uh” or “er”-type vowels with an “H” sound between them, usually = disagreement; OR, a sound like the word “ahah!”; usually = success or realisation | uhu |

To limit the variety of forms entering the corpus, we instructed transcribers to map each sound they encountered to the most appropriate orthographic form from Table 5. For example, sounds that might loosely be represented by the spellings ‘mmm’, ‘mm-mm’ and ‘mm-hm’ are all to be captured by the form ‘mm’. The intended effect is maximized inter-transcriber consistency and, therefore, maximizing corpus query recall by conflating orthographic variants of very similar filled pause sounds.

- Non-linguistic vocalizations

Non-linguistic vocalizations were recorded within [square brackets]. To enhance consistency, we limited transcribers to a fixed list of vocalization types. We assessed the list of permissible non-linguistic vocalisations in the BNC1994 scheme. These include *cough*, *sneeze*, *laugh*, *yawn* and *whistling* (Crowdy 1994: 27). In addition to these, we added *gasp*, *sigh*

and *misc* – a miscellaneous category allowing transcribers to include any non-linguistic vocalisation which cannot be easily described. During pilot testing we encountered some instances of singing that caused difficulty for the transcribers. The easiest solution was simply to introduce the new vocalization tag '[sing=LYRICS]' (where LYRICS is replaced by the linguistic content of what is sung) to distinguish such cases from normal speech. This *sing* tag also accounts for instances of unclear but yet tuneful speech – i.e. the singing of a melody with no words. In these cases, LYRICS is replaced by a question mark.

4.3.3 Transcription procedures and quality control

Multiple procedures were incorporated into the transcription process to ensure the consistent and accurate application of the transcription scheme described above, and thus, to apply quality control to the final corpus.

As audio files were received by the CUP team from contributors, checks were conducted to ensure that their quality was clear enough for good orthographic transcription. These checks involved listening to samples of each audio files and assessing its quality; the best files had a clear audio signal with minimal disruptive background noise. Audio files which passed the checks were sent in batches to a team of twenty transcribers, who had each been trained to use the transcription scheme. The transcribers were in regular contact with the CUP team to discuss and clarify any areas of uncertainty, and they were able to reject further audio recordings if they discovered any previously undetected quality issues.

The transcribers used Microsoft Word to transcribe the audio files. This meant that they were required to type each tag manually whenever it was required. While this created a risk of typing errors, the combination of systematic quality control measures followed by automated error detection and correction (see below) meant that no such errors entered the XML files that were the final product.

After each recording was transcribed, the transcript was put through two stages of checking – audio-checking and proofreading. At the audio-checking stage, a randomly-selected 5% sample of the audio recording was checked against the transcript for linguistic accuracy. If errors were found, the entire recording was checked. After this, the entire transcript was proofread for errors with regard to the transcription conventions (without reference to the audio).

Despite this checking, complete accuracy and consistency of transcription cannot, of course, be assumed. It is unavoidable that the involvement of twenty human transcribers will lead to certain inconsistencies of transcription decisions. Our extended and elaborated transcription scheme enabled us to minimize – but not to eradicate – such inconsistency. We encourage users not to consider the data 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 believe to be the natural, terminal limit of consistency between human transcribers. Furthermore, we explicitly facilitate the exploration of possible inter-transcriber inconsistency by including ‘transcriber code’ as a text metadata category (see Section 4.2.2).

4.3.4 Speaker identification

Another important aspect of spoken corpus transcription has no bearing on the accuracy of the transcription of linguistic content itself (i.e. what was said), but relates to the identification of the speaker that produced the transcribed turn (i.e. who said it) – in other words, how accurately, and with what degree of confidence, the transcribers were able to identify the speaker responsible for each turn. This process of ‘speaker identification’ was recorded in the transcripts by way of speaker ID codes, which are unique to each individual speaker in the corpus:

<0211> I haven’t met you

<0216> oh hi

The above example – shown in transcription format rather than canonical XML – demonstrates how two speakers, in this case 0211 and 0216, are distinguished in the transcript. The global uniqueness of the codes is crucial for the organization of the corpus according to categories of speaker metadata, since each code corresponds to the metadata of an individual speaker in the corpus.

When the transcribers assigned a code to a turn, they had three options available to indicate their confidence of their selection of a given speaker:

(1) CERTAIN

- mark the turn using a speaker ID code (e.g. <0211>); or,

(2) BEST GUESS

- mark the turn using their 'best guess', with a question mark added to the speaker ID code (e.g. <0211?>); or,

(3) INDETERMINABLE

- mark the turn only according to the gender of the speaker (i.e. <M> or <F>) or show that many speakers produced a turn (i.e. <MANY>).

The 'certain' codes occur when the transcribers selected an individual speaker as the producer of an individual turn; this is the most usual scenario. The 'best guess' code is intended for turns where the transcribers struggled to select an individual speaker with certainty, but felt able to provide a less confident 'best guess'. 'Indeterminable' codes occur when the transcribers were so uncertain that they were unable to provide a 'best guess', but could at least specify the gender of the voice they heard.

Precision of speaker identification – the accuracy with which speaker ID codes are assigned to turns in a transcript – is a previously unexplored issue in spoken corpus development. Thorough investigation into speaker identification revealed that the Spoken BNC2014 transcribers gave speaker IDs with high levels of confidence, even for recordings containing high numbers of speakers. However, inter-rater agreement and accuracy for such recordings is relatively low – low enough that, should these texts be used for sociolinguistic purposes, researchers run a reasonable risk of observing effects which are caused not by true language variation but by erroneously-identified speakers. That said, when an incorrect speaker ID was given, the incorrect speaker was almost always of the correct gender, and most of the time also in the correct age bracket(s). Another reassuring factor is that the vast majority of recordings feature only two or three speakers (such texts comprise three quarters of the corpus; see Table 6). We are confident that speaker identification in these less problematic transcripts is likely to have been conducted with acceptably high accuracy.

Table 6. Frequency of corpus texts per number of speakers per recording in the Spoken BNC2014.

| No. of speakers per recording | No. of Spoken BNC2014 texts | Cumulative percentage |
|-------------------------------|-----------------------------|-----------------------|
| two | 622 | 49.72 |
| three | 335 | 76.50 |
| four | 198 | 92.33 |
| five | 54 | 96.64 |
| six | 25 | 98.64 |
| seven | 11 | 99.52 |
| eight | 2 | 99.68 |
| nine | 3 | 99.92 |
| twelve | 1 | 100.00 |

We recommend that researchers should think carefully about whether to include texts which contain four or more speakers when conducting sociolinguistic research – especially work which looks at social groups (other than gender and age which, as shown, are likely to have been assigned with high levels of accuracy regardless of whether the precise speaker ID code was assigned correctly). To facilitate this, ‘number of speakers’ is a text metadata category (see Section 4.2.2) and is accessible within the restricted query function in CQPweb.

In addition, we made visible in the CQPweb interface the transcription convention for speaker identification confidence level. The purpose of this is to caution users against making the assumption that all of the speaker ID codes in the corpus texts have been assigned accurately. In the example below, for instance, the transcriber has indicated that they were not fully certain 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 Section 7.2).

S0511: well what happens in the sessions?

S0514[??]: there was some watching videos and stuff (SFQE)

Though this measure does not actually improve the accuracy of speaker identification, it does promote user awareness of the potential issues. Furthermore, this utterance-level attribute data makes it possible to restrict corpus queries to exclude turns with low

confidence in speaker identification (see Section 4.2.4). In total, 29,371 utterances (2.5% of utterances; 170,807 tokens) fall into the low confidence category.

5 The Written BNC2014

This section will be populated when the Written BNC2014 is made publicly available.

6 The BNC2014 on CQPweb

CQPweb (Hardie 2012) is an online corpus analysis system that acts as an interface to the Corpus Workbench software (CWB) and its powerful Corpus Query Processor (CQP) search utility. The architecture of CQPweb was based closely on that of an earlier tool, BNCweb (Hoffmann et al. 2008), which provided much of the same functionality, but only for the original BNC1994. By releasing the BNC2014 initially through Lancaster University's CQPweb server, we have, therefore, made it available first in a form in which many or most of the scholars who have worked on the earlier corpus will be familiar.

Many of the affordances of CQPweb have been referred to in earlier sections of the manual. This section summarises these and some other features of the system as applied to the BNC2014.

6.1 The Spoken BNC2014 in CQPweb

Concordance

Once you have logged in to the CQPweb server and entered into the interface of a particular corpus, the first thing you see is always a query entry screen (Figure 4). This is because CQPweb is organised around the assumption that most analyses will begin with a query of the corpus and the concordance display of the results found for that query. CQPweb supports two different query languages: the Simple query (which is the default), and CQP syntax (a more powerful and more formal query language which is also used in other software such as SketchEngine). Most users will begin with the simple query language and many never need to go beyond it.

Menu

Corpus queries

[Standard query](#)
[Restricted query](#)
[Word lookup](#)
[Frequency lists](#)
[Keywords](#)
[Analyse corpus](#)

Saved query data

[Query history](#)
[Saved queries](#)
[Categorised queries](#)
[Upload a query](#)
[Create/edit subcorpora](#)

Corpus info

[View corpus metadata](#)
[Corpus documentation](#)
[Oxford Simplified Tagset](#)
[CE tagset](#)
[USAS tagset](#)

About CQPweb

[CQPweb main menu](#)
[Your user page](#)
[Open help system](#)
[Video tutorials](#)
[Who did IT?](#)
[Latest news](#)
[Report bugs](#)

Spoken BNC2014: powered by CQPweb

Standard Query

Query mode: Simple query (ignore case) [Simple query language syntax](#)

Number of hits per page: 50

Restriction: None (search whole corpus)

[Start Query](#) [Reset Query](#)

System messages

BNC unrestricted access

2016-10-16

Following a cue from [bncweb.lancs.ac.uk](#) I have now made access at normal level (ie. not "restricted") to the BNC (XML-edition) available to everybody.
Andrew.

Restricted access to more corpora

2015-10-15

As of October 2015 I have implemented "restricted access" - a level of permission that lets you work with a corpus, without being able to get at the full text.
I am therefore able to give EVERYONE access on this "restricted" basis to a set of corpora that were previously off-limits due to their licences.
Have fun!
Andrew.

Figure 4. Standard query screen for the Spoken BNC2014 in CQPweb (Hardie 2012).

When you run a query, the results are presented as a concordance – the classic search display also known as *Key Word In Context* (KWIC), showing a little of the text before and a little of the text after each hit (result) for the query (Figure 5). As well as the immediate context, the location of each hit in the corpus is shown (on the left of the display): the ID code for the text where the result occurs is shown, as a clickable link which leads through to a full view of all the metadata available for that text. In the Spoken BNC1014, moreover, an utterance number is given alongside the text ID, indicating how far into the text the utterance is in which the text occurs.

48

users. Another example is the underlying XML attribute-value pair *trans="overlap"*, which appears on the *<u>* (utterance) element to indicate that the beginning of the utterance overlaps in time with the end of the previous overlap (that is, an overlap transition: see Section 4.3.2). This is rendered in the interface as >>. The display format that we use for such features in CQPweb does not replicate the original codes as typed by the transcribers and described in 4.3.2; the display codes were instead devised afresh for maximal visual distinctiveness.

Restrictions and metadata

In the process of compiling the Spoken BNC2014, we collected a substantial amount of metadata about both the speakers in the corpus and the texts in the corpus (that is, the individual transcriptions and the recorded conversations from which they arise). As explained above, this metadata can be viewed via the concordance display. However, often you would want to actually use the metadata at the outset of the analysis, to limit the part of the corpus that is included in the initial query. For instance, you might be interested in the usage of a word specifically in the language of younger people (say, under-30s). Or you might wish to search only within texts with just two or three speakers. The way to do this is via a Restricted Query, which is available as an alternative option to the default Standard Query at the corpus's entry page in CQPweb (Figure 6).

Menu

Corpus queries

Standard query

Restricted query

Word lookup

Frequency lists

Keywords

Analyse corpus

Saved query data

Query history

Saved queries

Categorised queries

Upload a query

Create/edit subcorpora

Corpus info

View corpus metadata

Corpus documentation

Oxford Simplified Tagset

C6 tagset

USAS tagset

About CQPweb

CQPweb main menu

Your user page

Open help system

Video tutorials

Who did it?

Latest news

Report bugs

Spoken BNC2014: powered by CQPweb

Restricted Query

Query model:

Simple query (ignore case)

Simple query language syntax

Number of hits per page:

50

Start Query

Reset Query

Select the text-type restrictions for your query:

| Transcription conventions used | Sample release inclusion | Number of speakers |
|-----------------------------------|--|-----------------------------|
| <input type="checkbox"/> Original | <input type="checkbox"/> Texts not in Sample release | <input type="checkbox"/> 12 |
| <input type="checkbox"/> Revised | <input type="checkbox"/> Texts within Sample release | <input type="checkbox"/> 2 |
| | | <input type="checkbox"/> 3 |
| | | <input type="checkbox"/> 4 |
| | | <input type="checkbox"/> 5 |
| | | <input type="checkbox"/> 6 |
| | | <input type="checkbox"/> 7 |
| | | <input type="checkbox"/> 8 |
| | | <input type="checkbox"/> 9 |

| Recording period | Year of recording | Transcriber |
|----------------------------------|-------------------------------|------------------------------|
| <input type="checkbox"/> 2012 Q1 | <input type="checkbox"/> 2012 | <input type="checkbox"/> T01 |
| <input type="checkbox"/> 2012 Q2 | <input type="checkbox"/> 2013 | <input type="checkbox"/> T02 |
| <input type="checkbox"/> 2013 Q3 | <input type="checkbox"/> 2014 | <input type="checkbox"/> T03 |
| <input type="checkbox"/> 2014 Q1 | <input type="checkbox"/> 2015 | <input type="checkbox"/> T04 |
| <input type="checkbox"/> 2014 Q2 | <input type="checkbox"/> 2016 | <input type="checkbox"/> T05 |
| <input type="checkbox"/> 2014 Q3 | | <input type="checkbox"/> T06 |
| <input type="checkbox"/> 2014 Q4 | | <input type="checkbox"/> T07 |
| <input type="checkbox"/> 2015 Q1 | | <input type="checkbox"/> T08 |
| <input type="checkbox"/> 2015 Q2 | | <input type="checkbox"/> T09 |
| <input type="checkbox"/> 2015 Q3 | | <input type="checkbox"/> T10 |
| <input type="checkbox"/> 2015 Q4 | | <input type="checkbox"/> T11 |
| <input type="checkbox"/> 2016 Q1 | | <input type="checkbox"/> T12 |
| <input type="checkbox"/> 2016 Q2 | | <input type="checkbox"/> T13 |

Figure 6. Restricted Query screen in CQPweb.

The Restricted Query presents the same basic search tool as the Standard Query, but in addition, there are several extra tables laying out different categories of texts or speakers. The categories that restrict a query to certain types of text are directly below the query controls; the categories concerning types of speaker are found further down on the page. If you select a category before running a query, then CQPweb will only search within that category. You can select more than one type of restriction at once (e.g. you could combine restrictions based on number of speakers in the text, year of recording, and age of speaker). If you don't select any categories of a particular type, then that criterion is simply not used: e.g. if you select neither *Male*, nor *Female*, nor *N/A* from the list of gender categories, then the query will retrieve results from utterances by speakers of any gender.

Collocation

Once you have run an initial query, you will in many cases want to apply further analyses to reorganise or summarise the concordance data. CQPweb has several tools for follow-up analysis of a concordance – all accessed from the concordance display's control menu.

There is not space to deal with all of them here, so we will focus on just the Collocation and Distribution tools.

Generally speaking, a collocation analysis looks for items (words or tags) in the co-text of some node item. Items which are highly frequent, or more frequent than expected, in the vicinity of the node are described as that node's collocates. In CQPweb, the collocation system builds a list of nearby items from the results in the concordance – the 'node' is simply whatever you searched for, whether that's an individual word, a phrase, or something more complicated. CQPweb then allows you to explore this data using different statistical measures and other tweaks to the method.

In practical terms, this means that when you enter the Collocation tool the first thing you have to do is choose the settings for the database of nearby items around the node. For all but advanced uses, the default settings are normally fine (Figure 7).

| Choose settings for proximity-based collocations: | | | |
|---|--------------------|--|--|
| Include annotation: | Simple POS | <input type="radio"/> Include | <input checked="" type="radio"/> Exclude |
| | Full USAS analysis | <input type="radio"/> Include | <input checked="" type="radio"/> Exclude |
| | Lemma | <input type="radio"/> Include | <input checked="" type="radio"/> Exclude |
| | Part-of-speech tag | <input checked="" type="radio"/> Include | <input type="radio"/> Exclude |
| | Semantic tag | <input type="radio"/> Include | <input checked="" type="radio"/> Exclude |
| | Tagged lemma | <input type="radio"/> Include | <input checked="" type="radio"/> Exclude |
| Maximum window span: | + / - 5 ▾ | | |
| Create collocation database | | | |

Figure 7. Collocation settings control.

Once the data is collected, you are shown the Collocation display (Figure 8). At the top of this display are controls that allow you to adjust how the collocates are calculated, with the actual table of collocates, ordered by their statistical score and presented with additional quantitative information, shown in the rest of the display below the controls.



Figure 8. Collocates of the node 'love', ranked by log-likelihood.

Distribution

The Distribution tool makes use of the corpus's metadata, just like a Restricted Query does. The difference is that the Distribution display shows you differences in the frequency – both absolute and relative – of the word or phrase you have searched for across different sections of the corpus.

By default, you are shown the distribution of your results across categories of texts (Figure 9). However, since the Spoken BNC2014 also contains speaker metadata (see Section 4.2) you can switch to viewing distribution across the categories of speaker (age, gender, and so on – Figure 10). Both can be displayed either as tables or as bar charts. An interesting additional function is the tool to look at text-frequency or speaker-frequency extremes. This means that, once you have searched for a word, you can find out the ID codes of the speakers who use that word (relatively) most or least often.



Figure 9. Distribution of search term 'love' across categories of texts, viewed as tables.



Figure 10. Distribution of search term 'love' across categories of speakers, viewed as bar charts.

Keywords and subcorpora

An alternative ‘way in’ to a corpus, rather than starting with a query for a word or phrase that you already know is of interest to you, is to use quantitative techniques to identify words that are likely to be of interest because of their unusually high frequency – relative to some comparison point. This technique is called a keywords analysis (or more generally, key items, since it can be run on tags as well as words) and can be seen as a more exploratory, or bottom-up, approach than beginning with a single search.

The Keywords tool in CQPweb (Figure 11) is another option accessible from the corpus entry screen. Once you access it, you can select two corpus frequency lists to compare, for instance, the Spoken BNC2014 itself as corpus 1, and some other English dataset as corpus 2. (Once it is complete the Written BNC2014 will make an obvious point of comparison!)

The screenshot shows the CQPweb interface for the Keywords tool. On the left is a sidebar menu with categories: Menu, Corpus queries (Standard query, Restricted query, Word lookup, Frequency lists, Keywords, Analyse corpus), Saved query data (Query history, Saved queries, Categorized queries, Upload a query, Create/edit subcorpora), Corpus info (View corpus metadata, Corpus documentation, Oxford Simplified Tagset, CB tagset, USAS tagset), and About CQPweb. The main area is titled 'Spoken BNC2014: powered by CQPweb' and 'Keywords and key tags'. It explains that keyword lists are compiled by comparing frequency lists. There are two 'Select frequency list' dropdowns, both set to 'Whole of Spoken BNC2014'. A 'Compare' dropdown is set to 'Word forms'. Below is the 'Options for keyword analysis' section, which includes a 'Show' dropdown set to 'All keywords', a 'Comparison statistic' dropdown set to 'Loglikelihood', a 'Significance cut-off point' dropdown set to '0.01%', a checked 'Use Sidaa correction?' checkbox, and two 'Min. frequency' input fields, both set to '3'. A 'Calculate keywords' button is located below these options. At the bottom, there is a section 'View unique words or tags on one frequency list:' with a dropdown set to 'Frequency list 1 but NOT frequency list 2' and a 'Show unique items on list' button.

Figure 11. Keywords screen.

As with Collocation, CQPweb’s Keywords system offers many options to tweak the details of the statistical procedure that will be used. Whatever you choose, running the analysis will lead to a list of words that are distinctive of corpus 1 versus corpus 2 (and/or vice versa) – ordered by their statistical score. However, CQPweb is still designed to emphasise the importance of looking at, and interpreting, words and other items in their actual context – so each entry in the Keywords table is actually a link through to a query for that word.

You can also use the Keywords tool to compare different sections of the Spoken BNC2014 to one another. To do this, you first have to define the sections for use in the comparison as subcorpora. This is yet another option accessible from the corpus entry screen. There are various methods for creating a subcorpus, but the most common is defining a subcorpus using corpus metadata – that is, using the same category-selection controls as in the Restricted Query function. Once you have created a subcorpus and compiled its frequency list, it will be available in the Keywords tool for use in comparisons.

| Spoken BNC2014: powered by CQPweb | | | | | | |
|---|---------------------------|---------------|----------------|------------------|--------|--|
| Create and edit subcorpora | | | | | | |
| Define new subcorpus via: Corpus metadata Get | | | | | | |
| Existing subcorpora | | | | | | |
| Name of subcorpus | Size | Size in words | Frequency list | Actions | Delete | |
| Last restrictions | 29,371 Utterance units | 170,807 | N/A | [copy] [add] [x] | | |
| 0_22 | 515,920 Utterance units | 5,033,510 | Available | [copy] [add] [x] | | |
| 1998_age | 1,075,720 Utterance units | 10,129,079 | Compile | [copy] [add] [x] | | |
| 60_above | 210,196 Utterance units | 1,845,576 | Available | [copy] [add] [x] | | |
| Cafe | 10 texts | 106,686 | Available | [copy] [add] [x] | | |
| NORTH_ENGLAND | 225,138 Utterance units | 2,208,480 | Available | [copy] [add] [x] | | |
| SOUTH_ENGLAND | 529,606 Utterance units | 4,982,755 | Available | [copy] [add] [x] | | |
| Sydney_restaurant | 1 text | 9,690 | Available | [copy] [add] [x] | | |

Figure 12. Create/edit subcorpora screen.

Advanced features

CQPweb also supports many more specialised analyses and procedures – including annotation of saved concordances; exploratory statistical analysis and visualisation; and thinning, randomisation and reduction of query result sets. While there is no space here to explore all these functions, many of them are explained by CQPweb’s built-in video help files (available by clicking “video tutorials” in the side menu). We hope you find these systems valuable and a source of insight in your exploration of the Spoken BNC2014.

6.2 The Written BNC2014 in CQPweb

This section will be populated when the Written BNC2014 is made publicly available.

7 Encoding and markup

7.1 Overall XML structure

As already noted, the design of the XML markup of the BNC2014 closely follows the recommendations of Hardie (2014). Each text in the corpus is stored within a separate XML file (in UTF-8 text encoding). All files in the corpus have the following overall structure:

```
<text id="XXXX">
  <header>
    (...metadata goes here...)
  </header>
  <body>
    (...actual corpus text goes here...)
  </body>
</text>
```

The unique text identification code is given within in the *id* attribute of the `<text>` element (i.e. in place of XXXX in the example above). All text ID codes are alphanumeric and begin with a letter. Texts in the spoken corpus have IDs beginning in ‘S’. Texts in the written corpus have IDs beginning in any other letter. Other than this, the text ID codes are arbitrary. The `<text>` element always contains exactly one `<header>` element followed by exactly one `<body>` element. The former contains a structured XML representation of the text metadata, and the latter contains the text itself. The precise list of tags used in both the `<header>` and the `<body>` differ between the spoken and written corpora; a complete Document Type Definition (DTD) is provided for the spoken corpus in its XML-format release, and is reproduced in this manual in Appendix D; a similar DTD for the written corpus will be provided upon its XML-format release.

7.2 Spoken corpus document XML

The body of each spoken corpus document consists of a transcript created according to the procedures described in Section 4.3.3, and each XML element used reflects one of the conventions described in Section 4.3.2. Where possible we limited ourselves to elements and attributes noted by Hardie (2014: 94-101) as having become more-or-less established as

de facto standard for spoken corpora – most of which are in fact also part of TEI and CES; we made additions to this set of codes only where our transcription scheme required it. For instance, utterances are marked up with `<u>` tags, and each utterance has a *who* attribute, containing the unique ID code of one of the 668 speakers (or one of three special codes indicating a male, female or multiple unknown speaker). Moreover, the use of an *n* attribute on `<u>` (indicating the sequential position of each utterance in its text) is also a *de facto* standard. These are all 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 identification (see Section 4.3.4). The default value of *whoConfidence* is *high*; it is only given explicitly when its value is *low*. Likewise, we added a *trans* attribute indicating the utterance transition type: the default value of this is *nonoverlap* and it is present explicitly when its value is *overlap*.

Table 7 (overleaf) shows how each transcription convention was converted into the Spoken BNC2014 XML.

Figure 13, which follows, presents a line-by-line comparison of an excerpt from a corpus text in both the original format typed by the transcribers (left column), and the canonical XML format into which it was converted (right column). This demonstrates some of the features of the corpus XML. As mentioned, each utterance is enclosed by `<u>` tags, with attributes for utterance number (*n*), speaker ID code (*who*) and, where relevant, confidence (*whoConfidence*). Line 3 of the XML includes examples of both versions of the `<unclear>` tag: ‘unclear word, guessed by transcriber’ (`<unclear>into</unclear>`), and ‘unclear word, no guess’ (`<unclear/>`). Finally, line 8 shows how the de-identification tags appear in XML in the form of an `<anon>` element containing an attribute for *type* (whose value is, in this case, *place*, indicating that the name of a place has been omitted).

Table 7. The main tags from the Spoken BNC2014 transcriptions scheme in both conventional and XML format

| Feature | Transcription scheme | XML |
|----------------------------|----------------------|---|
| speaker ID | <001> | <u who="S0001"> |
| uncertain speaker ID | <001?> | <u who="S0001" whoConfidence="low"> |
| male speaker | <M> | <u who="UNKMALE" whoConfidence="low"> |
| female speaker | <F> | <u who="UNKFEMALE" whoConfidence="low"> |
| multiple speakers | <MANY> | <u who="UNKMULTI" whoConfidence="low"> |
| anonymized male name | <name M> | <anon type="name" nameType="m" /> |
| anonymized female name | <name F> | <anon type="name" nameType="f" /> |
| anonymized neutral name | <name N> | <anon type="name" nameType="n" /> |
| anonymized place | <place> | <anon type="place" /> |
| telephone number | <tel-num> | <anon type="telephoneNumber" /> |
| address | <address> | <anon type="address" /> |
| email address | <email> | <anon type="email" /> |
| bank details | <bank-num> | <anon type="financialDetails" /> |
| social media username | <soc-med> | <anon type="socialMediaName" /> |
| date of birth | <DOB> | <anon type="dateOfBirth" /> |
| other personal information | <pers-inf> | <anon type="miscPersonalInfo" /> |
| false starts and repairs | = | <trunc>material-before</trunc> |
| truncated words | = (.) | No separate translation, just uses the normal combination of truncation plus pause. |
| overlap | | Adds <i>trans="overlap"</i> to the preceding <u> tag. |
| guessed words | <u=GUESSEDWORDS> | <unclear>GUESSEDWORDS</unclear> |
| unintelligible speech | <u=?> | <unclear /> |
| laughter | [laugh] | <vocal desc="laugh" /> |
| cough | [cough] | <vocal desc="cough" /> |
| gasp | [gasp] | <vocal desc="gasp" /> |
| sneeze | [sneeze] | <vocal desc="sneeze" /> |

| | | |
|------------------------------------|---------------------|--|
| sigh | [sigh] | <vocal desc="sigh" /> |
| yawn | [yawn] | <vocal desc="yawn" /> |
| whistle | [whistle] | <vocal desc="whistle" /> |
| miscellaneous noise | [misc] | <vocal desc="misc" /> |
| singing | [sing=LYRICS] | <shift new="singing" />LYRICS<shift new="normal"/> |
| foreign languages | [f=LANGUAGE=WORDS] | <foreign lang="LANGUAGE">WORDS</foreign> note also, if "Language" is recognised it can be replaced by a standard 3-letter code from ISO-639-2 (e.g. fra, deu, spa); if ? is given, then it is lang="und" (for "undetermined") if no words given, then just <foreign lang="LANGUAGE" /> |
| nonsense / made-up words | [nonsense] | <vocal desc="nonsense" /> |
| short pause | (.) | <pause dur="short"/> |
| long pause | (...) | <pause dur="long"/> |
| background speech | [e=background talk] | <event desc="background talk" /> |
| unintelligible conversation | [e=unintelligible] | <event desc="unintelligible" /> |
| overlapping exchanges begin | [e=begin overlap] | <event desc="begin overlap" /> |
| overlapping exchanges end | [e=end overlap] | <event desc="end overlap" /> |
| sounds and noises | [e=sound of X] | <event desc="sound of X" /> |
| music | [e=music] | <event desc="music" /> |
| abrupt end of recording | [e=abrupt end] | <event desc="abrupt end" /> |
| people entering conversation venue | [e=S000I enters] | <event desc="S000I enters" /> |
| people leaving conversation venue | [e=S000I leaves] | <event desc="S000I leaves" /> |
| problems in recording | [e=recording skips] | <event desc="recording skips" /> |

[TEXT]

<0211> I haven't met you

<0216> oh hi

<0220> oh right okay I feel a bit weird about going <u=into> Geordie now but
<u=?>

<MANY> [laugh]

<0211> <u=?> or me

<MANY> [laugh]

<0216> what part of Newcastle are you from?

<0220> <place>

<0216> oh yeah

<0220> oh where you from?

<0216> <place>

<body>

<u n="1" who="S0211">I haven't met you</u>

<u n="2" who="S0216">oh hi</u>

<u n="3" who="S0220">oh right okay I feel a bit weird about going
<unclear>into</unclear> Geordie now but <unclear/></u>

<u n="4" who="UNKMULTI" whoConfidence="low"><vocal desc="laugh"/></u>

<u n="5" who="S0211"><unclear/> or me</u>

<u n="6" who="UNKMULTI" whoConfidence="low"><vocal desc="laugh"/></u>

<u n="7" who="S0216">what part of Newcastle are you from?</u>

<u n="8" who="S0220"><anon type="place"/></u>

<u n="9" who="S0216">oh yeah</u>

<u n="10" who="S0220">oh where you from?</u>

<u n="11" who="S0216"><anon type="place"/></u>

Figure 13. Transcript excerpt, pre- and post-XML conversion.

7.3 Spoken corpus header XML

The text headers in the spoken corpus use a much simpler (and more flatly organised) set of metadata tags than does the TEI. Each such header element was, in fact, generated automatically, on a mostly one-to-one basis, from some column of the metadata tables originally collected alongside the recordings. This metadata is listed and described elsewhere in this manual (see Sections 4.2.2 and 4.2.3).

The names of the elements within the header are the same as the short handles used for the equivalent metadata fields used in CQPweb (for the benefit of users transitioning from one mode of corpus access to another); explanations of all element names, and of codes used in some of the element values, are provided in textual format along with the XML files (in the sub-folder “metadata” within the distribution download bundle). Briefly, the content of a spoken corpus header is as follows:

- The `<header>` element contains the following series of elements: *rec_length*, *rec_date*, *rec_year*, *rec_period*, *n_speakers*, *list_speakers*, *rec_loc*, *relationships*, *topics*, *activity*, *conv_type*, *conventions*, *in_sample*, *speakerInfo*
- Each of these *except* `<speakerInfo>` contains a simple metadata item that corresponds to one of the items listed in sections 4.2.2 and 4.2.3.
- The `<speakerInfo>` element contains a series of `<speaker>` elements, each of which contains the metadata for one of the speakers included in the earlier `<speaker_list>`.
- Each `<speaker>` element, like the overall `<header>`, consists of a flat list of elements containing specific metadata items, as follows: *exactage*, *age1994*, *agerange*, *gender*, *nat*, *birthplace*, *birthcountry*, *l1*, *lingorig*, *dialect_rep*, *hab_city*, *hab_country*, *hab_dur*, *dialect_l1*, *dialect_l2*, *dialect_l3*, *dialect_l4*, *edqual*, *occupation*, *socgrade*, *nssec*, *l2*, *fls*, *in_core*.
- These elements correspond to the items of metadata described in sections 4.2.5 and 4.2.6.

Because some speakers appear in more than one corpus text, the metadata for those speakers will be repeated in the header of each text. For convenience, a full set of all speaker metadata in XML format is provided as file ***speakerInfo.xml*** alongside the XML download.

7.4 Written corpus document XML

This section will be populated when the Written BNC2014 is made publicly available.

7.5 Written corpus header XML

This section will be populated when the Written BNC2014 is made publicly available.

7.6 Changes made to the XML for annotation and CQPweb indexing

Certain automatic adjustments were made to the formatting of the XML prior to the texts being tagged and then indexed on Lancaster's CQPweb server. The purpose of these changes was to address certain requirements of the tagger software (see Section 8.1) and of CQPweb itself.

First, the headers were removed, as CQPweb requires metadata to be stored as a separate database rather than within the corpus itself; had the headers been left in, they would have been treated as a searchable part of the corpus text itself. With the header gone, there was no need for a distinct `<body>` tag, so this was removed as well, so that each `<text>` contains only the actual text of the corpus document. CQPweb requires the root element of each document to be `<text>`.

Secondly, all XML attributes with default values were made explicit. As noted above, the default values for *whoConfidence* (i.e. *high*) and *trans* (i.e. *nonoverlap*) were not actually included in the canonical XML (as default values can be assumed); however, CQPweb does not support this feature of XML, and so the default values were explicitly inserted.

Third, it was necessary to re-represent certain empty XML elements as words, namely `<anon/>`, `<unclear/>` and `<foreign/>`. Each of these empty elements actually represents a word which was present in the audio recording, but not transcribed for one reason or another. In the canonical XML version, it sufficed to use an empty element to indicate the omitted material. However, this was not a suitable solution for CQPweb because all XML is treated as occurring *between* a pair of adjacent positions in the token stream – with only actual text *outside* the XML “taking up space” in the token stream. So, for instance, an empty `<anon/>` element between two words AAA and BBB would be treated as indicating that words AAA and BBB are *adjacent*, with the `<anon/>` tag representing a region that begins with the BBB token. This would result in an inaccurate representation of distance in the corpus – for instance, AAA and BBB would be treated by the system as a bigram, even though in reality they are separated by the word or words that were omitted for purposes

of de-identification. We determined to avoid this. (A secondary issue is that the CLAWS POS tagger relies on wordclass-to-wordclass transition probabilities for contextual disambiguation, and could easily be deceived by such a false bigram.)

This re-representation was accomplished by the insertion of dummy tokens: fake words that would not clash with any other words in the corpus at the point in the token stream where the empty element occurs. All these begin with two hyphens (to make sure they could not clash). Those that begin with --ANON represent different types of anonymization. Foreign and unclear words with no guessed transcription were converted to --FOREIGNWORD and --UNCLEARWORD respectively. In the latter two cases, the dummy tokens were enclosed within non-empty *<foreign>* and *<unclear>* elements – making them equivalently encoded to the already-present non-empty *<foreign>* and *<unclear>* elements already in the corpus. A list of all the dummy types, together with their frequencies and the tags they are assigned in POS tagging, is given in Table 8.

Table 8. Dummy word types in the Spoken BNC2014

| | Dummy word type | POS tag | Frequency |
|----|------------------------|---------|-----------|
| 1 | --ANONaddress | FU | 124 |
| 2 | --ANONdateOfBirth | FU | 21 |
| 3 | --ANONemail | FO | 36 |
| 4 | --ANONfinancialDetails | FU | 10 |
| 5 | --ANONmiscPersonalInfo | FU | 280 |
| 6 | --ANONnameF | NPI | 31246 |
| 7 | --ANONnameM | NPI | 36783 |
| 8 | --ANONnameN | NPI | 2614 |
| 9 | --ANONplace | NPI | 24914 |
| 10 | --ANONsocialMediaName | NPI | 88 |
| 11 | --ANONtelephoneNumber | FO | 55 |
| 12 | --FOREIGNWORD | FW | 465 |
| 13 | --UNCLEARWORD | FU | 77031 |

8 Annotation

The whole corpus is annotated for part-of-speech (POS), lemma and semantic category.

8.1 POS tagging

The POS tagging was conducted using the same systems as the original BNC1994 – most notably the **Constituent Likelihood Automatic Word-tagging System (CLAWS)**; Garside 1987). CLAWS is a **hybrid probabilistic/rule-based tagger**, which means that when a word may have more than one tag, depending on the context, the system uses both probabilistic modelling and a set of idiom-structure rules to try and identify the correct analysis in the present context. Thus, the system may output a single tag for a given token, or a list of tags with CLAWS' estimation of the probability of each being correct, expressed as percentages (see Garside & Smith 1997) – in which case it is usual practice to accept the most probable choice. The tagging process is summarised as follows (expanded slightly from Garside 1996: 173):

1. The input running text is read in, divided into individual tokens (including the separation of clitics such as *n't*, *'m*, *'re* etc.), and sentence breaks are recognised.
2. A list of possible tags is then assigned to each word, the main source being a lexicon.
3. A number of words in any text will not be found in the lexicon, and for these there is a sequence of rules to be applied in an attempt to assign a suitable list of potential tags.
4. Since the lists of potential tags from steps 2 and 3 are based solely on individual words, the next step uses several libraries of template patterns to allow modifications to be made to the lists of tags in the light of the immediate context in which the word occurs.
5. The next step is to calculate the probability of each potential sequence of tags, and to choose the sequence with the highest probability as the preferred one.
6. Finally the text and associated information about tag choice is output.

In a departure from the practice of the BNC1994, the BNC2014 uses the C6 tagset instead of the simpler C5 tagset.¹² C5 tags were used in order to achieve a simpler (and thus more

¹² Both tagsets are available on the CLAWS website: <http://ucrel.lancs.ac.uk/claws/>

reliable) system of POS tagging in the first release of the BNC1994 – the estimated error rate for POS tagging in the Spoken BNC1994 is 1.17%, which is only 0.03% higher than that of the Written BNC1994 (see Leech & Smith 2000). However, later BNC1994 releases use a parallel system of simple tags, or major word class tags, *alongside* the C5 tags (see Appendix G; we dub these the *Oxford Simplified Tagset* from its origin at Oxford University Computing Services). 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 classification 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.

A problem that arose in the tagging of the Spoken BNC2014 is that ‘standard’ POS tagger resources are usually based on written data, as is indeed the case for CLAWS. This may be problematic. If the tag probabilities used in the tagger’s statistical model do not match those of the type of text being tagged, then mistakes can be made by the tagger that would be avoided if its model had been trained on a matching kind of text. Systematic error can therefore be introduced by using tagger resources trained on written English for the analysis of spoken English data – the examples presented later in this section aim to demonstrate this point.

To address this problem, in the 1990s CLAWS was retrained for the tagging of the Spoken BNC1994; the set of spoken resources thus developed include “supplementary lexicons and lists of pattern templates for spoken data” (Garside 1995). Compilers of subsequent spoken corpora have taken a similar approach with success.

The spoken resources for CLAWS are still available. However, given their age, their performance on contemporary data was not known. For instance, some of the lexical POS probabilities may have shifted over time. We therefore undertook an exercise to assess whether the spoken resources would provide superior performance for the Spoken BNC2014 to the standard CLAWS resources for written language. We tagged one randomly selected Spoken BNC2014 text using both sets of resources, and compared the outputs. A comparison of the first 1,000 tagged tokens in this text between the two outputs suggests strongly that, despite the age of the data used to train the spoken version of tagger, it was able to tag the text with greater accuracy than the standard tagger. The written tagger’s error rate on these 1,000 tokens was 7.3% while the error rate of the spoken tagger was 2.5%. Of the 1,000 words, 67 were tagged differently by the two taggers. Of

these, 57 (85%) were found to have been tagged incorrectly by the written tagger but correctly by the spoken tagger, providing evidence that the spoken tagger resources tend to facilitate more accurate tagging decisions than the written tagger resources. The evidence derived was sufficient to justify the selection of the spoken resources over the written resources.

8.2 Lemmatisation

Lemmatisation was applied to the text at the same time as semantic tagging, i.e. taking the POS tagged text as input. The same lemmatiser was used as for the original BNC1994. Lemmas are given as all-lowercase, *even if* the word token itself has an initial uppercase letter or is all-uppercase. This even applies to proper nouns and other forms which would not normally be given in lowercase.

8.3 Semantic tagging

Semantic tagging was conducted using the UCREL semantic annotation system (USAS; Rayson et al. 2004). See Appendix F for the USAS tag set. In the corpus as indexed on Lancaster's CQPweb server, two different USAS annotations are present: one which includes only the most likely semantic tag for each token, and another which encodes all possible USAS analyses, as per the following example:

EXAMPLE TEXT:

I take regular breaks for red wine

SEMANTIC TAG on *take*:

A9

FULL USAS ANALYSIS on *take*:

|A9u|T1:3|C1|A1:1:1|M2|S7:1d|A2:1u|X2:4|S6u|S7:4u|N3|P1|M1|X2:5u|F1|F2|Q1:2|B3|

There are several things to note here. First, while the normal USAS tagset uses the full stop (.) as a separator within tags, and plus (+) and minus (-) to indicate polarity of a semantic category, here instead colon (:) is used as the separator, and *u* and *d* for plus and minus (mnemonic for *up* and *down*). This is to avoid clashes between the contents of the semantic tag and regular expression syntax. Second, the alternative analyses in the full annotation are separated by pipe characters (|): this allows CQPweb to treat this as a *feature set* (please see

the CQP tutorials on the Corpus Workbench website¹³ for details). Third, the “most likely tag only” annotation takes the first possible tag from the full analysis, and removes any modifiers such as *u* or *d* from it; this is in order to group together as many tokens as possible for statistical analysis (with no loss of information, since the fine grained ‘full’ analysis is still available).

8.4 XML for annotation

In the XML release of the corpus, a second copy of the corpus is provided which renders the annotated text as described here as XML. These files:

1. Contain no headers;
2. Incorporate the adjustments to the original XML described in section 7.6;
3. Wrap each token of the text within a `<w>` element.

The `<w>` elements have the following attributes, each representing a single annotation:

1. `pos` – part-of-speech tag (CLAWS C6: see 8.1)
2. `lemma` – lemmatiser output (see 8.2)
3. `class` – “simple” POS tag or major word-class (Oxford Simplified Tags: see 8.1)
4. `usas` – semantic tag (USAS tags: see 8.3)

For example:

```
<text id="S2AJ">
<u n="1" who="S0439" trans="nonoverlap" whoConfidence="high">
<w pos="RRQ" lemma="how" class="ADV" usas="Z5">how</w>
<w pos="VBDZ" lemma="be" class="VERB" usas="A3">was</w>
<w pos="PNI" lemma="everything" class="PRON" usas="Z8">everything</w>
<w pos="IW" lemma="with" class="PREP" usas="Z5">with</w>
<w pos="UH" lemma="erm" class="INTERJ" usas="Z4">erm</w>
<w pos="IW" lemma="with" class="PREP" usas="Z5">with</w>
<w pos="NPI" lemma="--anonnamen" class="SUBST" usas="Z99">--ANONnameN</w>
<w pos="DDI" lemma="this" class="ADJ" usas="M6">this</w>
```

¹³ Documentation for Corpus Workbench is available at <http://cwb.sourceforge.net/documentation.php> .

<w pos="NNTI" lemma="weekend" class="SUBST" usas="Tl:3">weekend</w>
<w pos="YQUE" lemma="PUNC" class="STOP" usas="">?</w>
</u>

References

- Andersen, G. (2016). Semi-lexical features in corpus transcription: Consistency, comparability, standardisation. *International Journal of Corpus Linguistics*, 21(3), 323-347.
- Atkins, A., Clear, J., & Ostler, N. (1992). Corpus Design Criteria. *Literary and Linguistic Computing*, 7(1), 1-16.
- Baker, P. (2010). *Sociolinguistics and Corpus Linguistics*. Edinburgh: Edinburgh University Press.
- Börjars, K., & Burridge, K. (2010). *Introducing English Grammar* (2nd ed.). London: Hodder Education.
- Bradley, H. (2013). *Gender* (2nd ed.). Cambridge: Polity Press.
- Burnard, L., & Bauman, S. (Eds.) (2013). TEI: P5 Guidelines. TEI Consortium. Retrieved from <http://www.tei-c.org/Guidelines/P5/> (last accessed June 2017).
- Collis, D. (2009). *Social Grade: A Classification Tool*. Retrieved from https://www.ipsos.com/sites/default/files/publication/6800-03/MediaCT_thoughtpiece_Social_Grade_July09_V3_WEB.pdf (last accessed September 2017).
- Crowdy, S. (1993). Spoken Corpus Design. *Literary and Linguistic Computing*, 8(4), 259-265.
- Crowdy, S. (1994). Spoken Corpus Transcription. *Literary and Linguistic Computing*, 9(1), 25-28.
- Dembry, C., & Love, R. (2014, October). *Spoken English in Today's Britain*. Cambridge Festival of Ideas, Cambridge University, UK. Available [here](#).
- Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., & Danae, P. (1993). Outline of discourse transcription. In J. A. Edwards, & M. D. Lampert (Eds.), *Talking Data: Transcription and Coding in Discourse Research* (pp. 45–89). Hillsdale, NJ: Lawrence Erlbaum.
- Gablasova, D., Brezina, V., McEnery, T., & Boyd, E. (2015). Epistemic Stance in Spoken L2 English: The Effect of Task and Speaker Style. *Applied Linguistics* 2015, 1-26.
- Garside, R. (1987). The CLAWS Word-tagging System. In R. Garside, G. Leech, & G. Sampson (Eds.), *The Computational Analysis of English: A Corpus-based Approach*. London: Longman.
- Garside, R. (1995). Using CLAWS to annotate the British National Corpus. Oxford: Oxford Text Archive. Retrieved from http://www.natcorp.ox.ac.uk/docs/garside_allc.html (last accessed August 2017).

- Garside, R. (1996). The robust tagging of unrestricted text: the BNC experience. In J. Thomas & M. Short (Eds.), *Using corpora for language research: Studies in the Honour of Geoffrey Leech* (pp. 167-180). Longman, London.
- 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.
- Hardie, A. (2012). CQPweb – Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, 17(3), 380-409.
- Hardie, A. (2014). Modest XML for corpora: not a standard, but a suggestion. *ICAME Journal*, 38, 73-103.
- Hasund, K. (1998). Protecting the innocent: The issue of informants' anonymity in the COLT corpus. In A. Renouf (Ed.), *Explorations in corpus linguistics* (pp. 13-28). Amsterdam: Rodopi.
- Hoffmann, S., Evert, S., Lee, D., & Ylva, B. (2008). *Corpus linguistic 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).
- Laws, J., Ryder, C., & Jaworska, S. (2017). A diachronic corpus-based study into the effects of age and gender on the usage patterns of verb-forming suffixation in spoken British English. *International Journal of Corpus Linguistics*, 22(3).
- Leech, G., & Smith, N. (2000). *Manual to accompany the British National Corpus (Version 2) with Improved Word-class Tagging*. Lancaster: UCREL. Retrieved from <http://ucrel.lancs.ac.uk/bnc2/bnc2error.htm> (last accessed August 2017).
- Love, R. (2015, November). *Spoken English in UK society*. ESRC Language Matters: Communication, Culture, and Society. International Anthony Burgess Foundation, Manchester, UK. Available [here](#).
- 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).
- MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

- McEnery, T., Love, R., & Brezina, V. (2017). Introduction: Compiling and analysing the Spoken British National Corpus 2014. *International Journal of Corpus Linguistics*, 22(3).
- NRS. (2014). *Social Grade*. Retrieved from <http://www.nrs.co.uk/nrs-print/lifestyle-and-classification-data/social-grade/> (last accessed September 2017).
- Nelson, G. (2002). International Corpus of English: Markup Manual for Spoken Texts. Retrieved from www.ice-corpora.net/ice/spoken.doc (last accessed September 2017).
- Payne, J. (1995). The COBUILD spoken corpus: Transcription conventions. In G. Leech, G. Myers, & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-up and Application* (pp. 203–207). Harlow: Longman.
- Rayson, P., Archer, D., Piao, S. L., & McEnery, T. (2004). The UCREL semantic analysis system. In M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)* (pp. 7-12). Paris: European Language Resources Association.
- Rose, D., & O'Reilly, K. (1998). *The ESRC Review of Government Social Classifications*. London & Swindon: Office for National Statistics & Economic and Social Research Council. Retrieved from <http://www.ons.gov.uk/ons/guide-method/classifications/archived-standard-classifications/soc-and-sec-archive/esrc-review/index.html> (last accessed September 2017).
- 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.
- Thompson, P. (2005). Spoken language corpora. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 59–70). Oxford: Oxbow Books.

List of appendices

| | |
|---|-----|
| Appendix A. The Spoken BNC2014 user licence | 74 |
| Appendix B. Spoken BNC2014 Frequently Asked Questions document..... | 77 |
| Appendix C. Spoken BNC2014 transcription scheme..... | 78 |
| Appendix D. DTD for the Spoken BNC2014..... | 103 |
| Appendix E. DTD for the Written BNC2014 | 107 |
| Appendix F. UCREL CLAWS6 tagset (http://ucrel.lancs.ac.uk/claws6tags.html)..... | 108 |
| Appendix G. The Oxford Simplified Tagset (adopted from http://www.natcorp.ox.ac.uk/docs/URG/codes.html#klettpos) | 112 |

Appendices

Appendix A. The Spoken BNC2014 user licence

Preamble

The Spoken BNC2014 is a publicly-accessible resource. This means that anyone may obtain a copy and use it for non-commercial research. However, the materials contained within the corpus **are not** in the public domain. The copyright is owned by Cambridge University Press.

Use of the Spoken BNC2014 without registering and submitting a request via this form is forbidden. By registering and submitting this form you are entering into a license with Cambridge University Press. In using the Spoken BNC2014 you are bound by the following terms and conditions.

Terms used in this licence

- **The Corpus:** the Spoken British National Corpus 2014, including (a) the texts of the Corpus, (b) any modified versions of this Corpus supplied alongside those texts, and (c) all supplementary documentation and other material supplied alongside those texts.
- **We/Us:** Lancaster University, distributor of the Corpus, acting on our own behalf and on behalf of our partner Cambridge University Press as copyright holders in the material contained within the Corpus.
- **You:** the signatory of this licence, to whom permission to access and use the Corpus is granted.
- You may sign this licence either as an **individual**, or as a representative of an **institution**. In cases where different conditions apply to individual and institutional signatories, this is stated explicitly below: “If you are an **individual signatory**...” / “If you are an **institutional signatory**...”

General use of the Corpus

- **You may** make use of the Corpus only: (a) for purposes of non-commercial research, or (b) for purposes of teaching.
- **You must** at all times respect the privacy of speakers in the texts of the Corpus. In particular, **you must not** attempt to undo or bypass any of the anonymisation in any of the texts in the Corpus.
- If you are an **individual signatory**, **you must** ensure that your use of the Corpus, or use of the Corpus by other persons through any interface created or maintained by you, adheres to the terms of this licence.
- If you are an **institutional signatory**, **you must** ensure that use of the Corpus by any person affiliated to your institution, or who gains access to the Corpus through an interface created or maintained by your institution, adheres to the terms of this licence.

Publication of research

- **You may** publish the results of research that uses the Corpus.
- In any such publication, **you may** reproduce excerpts of the texts of the Corpus only as permitted under UK copyright law and “fair dealing”. **You must** clearly identify any such excerpt as originating from the Corpus and as owned by Cambridge University Press.
- In any such publication, **you must** acknowledge the use of the Corpus in your research, by citing the following standard reference (in your field’s usual referencing style):

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

- **We ask you to (but you do not have to)** inform us of such publications through our website, so that we can list them in appropriate public bibliographies of research that uses the Corpus.

Reproduction and modification of the Corpus

- If you are an **individual signatory**, **you may** make an unlimited number of copies of the Corpus for your personal use only.
- If you are an **institutional signatory**, **you may** make an unlimited number of copies of the Corpus for use by people affiliated to your institution (that is: employees and, if you are an educational institution, students). Likewise, **you may** copy the Corpus to a shared drive or network location within your institution, so long as this location can only be accessed by people affiliated to your institution.
- **You must not** redistribute the Corpus. This means that **you must not** transfer, or allow to be transferred, any copy (in part or full) of the Corpus or a modified version of the Corpus, to any other person or institution.
- **You must not** allow any other person or institution to access or use the Corpus, except under the conditions outlined below for online interfaces.
- **You must** store all your copies of the Corpus on computer equipment that you own and is under your direct control. In particular, **you must not** store any copy of the Corpus on any external “Cloud” Internet service.
- **You may** re-encode, reformat, annotate, and/or modify the Corpus in any way, and make use of, and/or make copies of, such a modified version of the Corpus in any of the ways that this licence permits.
- **You must not** pass copies of any such modified version of the Corpus, or any part of such a modified version, to any other person or institution (if you are an institutional signatory: to any person not affiliated to your institution as defined above).
- If you wish to allow others to obtain a copy of such a modified version of the Corpus, **you may** submit the modified version to us for distribution alongside the original Corpus.
- **We** reserve the right to accept or reject any such submission. If we accept such a submission, we reserve the right to distribute the modified version under a licence with more restrictive terms than those outlined here.
- **You** hereby **agree** that any intellectual property that you hold in a modified version of the Corpus, that is submitted to us and that we accept, shall be assigned to Cambridge University Press.

Use in online interfaces

- **You may** allow others to make use of your copy of the Corpus, or any modified version, via an online interface.
- **You must** ensure that any such online interface allows the Corpus to be used only in accordance with this licence. In particular:
- **You must** provide us, on request, with access to any such online interface.
- **You must** provide us, on request, with full details in writing of how access to the corpus data in any such online interface is monitored and controlled in accordance with the conditions above.

Commercial use of the Corpus

- **You must not** make use of the Corpus for any commercial or profit-making purpose under this licence.
- **You may** apply separately to us in writing for permission to make commercial use of the Corpus under a commercial licence.
- **We** reserve the right to refuse such requests, or to grant them subject to payment and/or subject to a licence with more restrictive terms than those outlined here.

Other conditions of use

- **You** hereby **acknowledge** that the Corpus data is provided to you “as is”, without any warranty, without even the implied warranty of fitness for a particular purpose.
- **You** hereby **agree** that we will not be liable to you for loss of profits, goodwill or any kind of consequential losses of any nature arising from your use of the Corpus, even if such loss was foreseeable.

Termination of the licence

- **We may** terminate this licence at any point, by giving you notice in writing. **You must** erase all copies of the Corpus in your possession upon receipt of such notice.
- **You may** terminate this licence at any point, by erasing all copies of the Corpus in your possession.

Data Protection

We will use your data in accordance with the Data Protection Act.

Appendix B. Spoken BNC2014 Frequently Asked Questions document



British National Corpus 2014 - Frequently Asked Questions

1. How long will the project be running for?

This is a long term project, that will run until at least June 2015, possibly longer.

2. How long do my recordings need to be?

There is no set length that your recordings need to be. You will be paid per hour of good quality recording that you submit, but the length of the recordings themselves is completely up to you.

For instance, you may choose to send in a few long (e.g. 2 hour) recordings or a series of shorter (e.g. 10 minute) recordings.

3. What should I talk about in my recordings?

We are looking for examples of natural conversation, so you can talk about anything you like. You might find it helpful to decide a few topics to cover in order to get the conversation started, but it is important that the discourse is natural and not too prescribed.

4. Do I need to use professional recording equipment?

No, professional recording equipment is not necessary. All we ask is that you produce a digital recording that can be sent to us in mp3 format. It is also important that the quality is good as we will not be able to pay you for recordings that are unclear or inaudible. Please contact us if you have any questions about this.

5. In what format should I send my recordings?

All recordings should be sent to us in an mp3 format. If you need help converting your recordings to this format, please email corpus@cambridge.org and we will send you some instructions.



6. Can I fill in all my forms electronically?

Most forms can be filled in electronically. The only exceptions are your signed contract and your signed speaker consent forms. Both of these will need to be posted to us as soon as you have completed them.

7. Do I need to send in hard copies of the speaker consent forms?

We will need both scanned copies of your speaker consent forms when you receive them, and hard copies which can be sent to us at the end of the project.

8. How do I share my recordings through Dropbox?

The easiest way to get your recordings to us is for you to create a Dropbox folder in which you save all your recordings. You can then share this folder with us (corpus@cambridge.org). If you need any help with this, please email us and we will send you a set of instructions.

9. Will I receive a contract?

When you have sent us your postal address and contact telephone number, we will send out a contract for you to sign.

You can then return the signed contract to us. In order to speed up the process, we also ask that you send us a scanned copy of your signed contract, if possible.

10. How do I get paid for the work I have submitted?

You can be paid for all your submissions at the end of each month. When you have submitted all the recordings you wish to, we will ask you to fill out a form for your bank details along with an invoice. These can both be filled in electronically.

Please note that we will not be able to pay you unless we have received all the necessary documents including recording information sheets for each recording, speaker consent forms for each participant, and your signed contract.

Appendix C. Spoken BNC2014 transcription scheme

Cambridge University Press – Transcription Conventions 5.0

Transcription conventions are used to indicate features of a spoken interaction (such as speaker turns, repetition and overlaps) in typed text. This document outlines the format of the conventions used by Cambridge University Press. This document should be used for reference throughout your transcription work for Cambridge.

The conventions are, in many respects, a working document (and you may receive an update from time to time). We would very much appreciate your comments in order to explain and refine our definitions further - please contact corpus@cambridge.org with any suggestions for future updates.

| CONTENTS | |
|-----------------------------------|---|
| <u>1. General guidelines</u> | <u>11. Unfinished words</u> |
| <u>2. Document format</u> | <u>12. Overlaps</u> |
| <u>3. Line height and spacing</u> | <u>13. Unintelligible speech</u> |
| <u>4. Header information</u> | <u>14. Acronyms, spelling and capitalisation</u> |
| <u>5. Tag format</u> | <u>15. Numbers</u> |
| <u>6. Speaker IDs</u> | <u>16. Non-standard words or sounds</u> |
| <u>7. Anonymization</u> | <u>17. Non-standard contractions or shortenings</u> |
| <u>8. Utterances</u> | <u>18. Speaker accent/dialect</u> |
| <u>9. Punctuation</u> | <u>19. Non-linguistic vocalisations</u> |
| <u>10. Pauses</u> | <u>20. Events</u> |

1. GENERAL GUIDELINES

Unlike legal or medical audio transcription, transcription for linguistic research requires you to transcribe exactly what you hear. This means that you should not correct or paraphrase any instances of “bad” grammar, unfinished sentences, missing or repeated words. We are very interested in this type of variation, so it is important that the transcription is a direct copy of the recording.

On the whole, accent features (i.e. the sounds of the language) should not be represented in the transcriptions. For example, speakers with regional or international accents may pronounce a word in a way that is different to what you might expect to find in Standard English, but no effort should be made to reflect this in the transcription. (This is discussed further in 18. SPEAKER ACCENT/DIALECT).

2. DOCUMENT FORMAT

When undertaking a transcription project for Cambridge, you will be sent a template file. You should open this template and then save your new file using the same name as the sound file you're working on. The name of the sound file and the text file should correspond.

For example, if you were transcribing the sound file *001.002.mp3* you should open the template and then save a copy of this file and call it *001.002.doc*.

You should use the same template for all transcriptions you work on. You should not change the font, spacing, justification, margins or anything else in this document.

3. LINE HEIGHT AND SPACING

Single line height should be used throughout your transcription. A double carriage return (enter key) should be used after each speaker as shown in 6. SPEAKER IDs.

4. HEADER INFORMATION

Header information is used to make a note of certain characteristics of the spoken data file. The following information (or similar) is saved in the template that you will work from and will appear as header information at the top of each transcription;

[HEADER]

FILE NAME:

MAIN SUBJECT:

LIST OF SPEAKERS:

[TEXT]

Please **do not delete** the [HEADER] and [TEXT] lines; these indicate where the header begins and ends.

When we send you a soundfile, we will also send you the list of speakers that feature in the recording. It will be a list of the speaker IDs in order, separated by commas, with no full stop at the end. Please copy and paste this into the header of the transcription. This then needs to be checked – see 6. SPEAKER IDS.

Copy the list of topics covered in the conversation from question 9 of the Recording Information Sheet and add it to MAIN SUBJECT

5. TAG FORMAT

Tags form the basis of the transcription conventions and are used to indicate features such as speaker turns and overlaps. Most tags have angle brackets < >. Some use other brackets (to make it easier to tell them apart).

Each tag also has a label to differentiate them from each other. E.g. is the tag for overlaps. These tags are explained in more detail in the subsequent sections¹⁴.

¹⁴ (NB – tags and certain words are shown in red throughout this document for emphasis and clarity only – all text and tags should be in transcribed black, as standard)

6. SPEAKER IDS

At the beginning of a new project you will be sent a spreadsheet containing information about all of the speakers in that project. You will be given their name, gender, first language and accent/dialect, plus a unique speaker ID number. These are in the format <001>, <002>, etc.

When you are sent a new file to transcribe, it will be accompanied by information about the recording – speaker names, their first utterance, plus other additional information.

For example, the recording information sheet you receive may contain the following:

| |
|---|
| Speaker 1: Anna Brown, 'OK, so are we recording now?' |
| Speaker 2: Thomas Brown, 'Yep' |

These names can be matched up with the speaker numbers using the spreadsheet. For example:

| Speaker ID | Name | Age | Gender | LI | Accent/dialect |
|-------------------|--------------|------------|---------------|-----------|-----------------------|
| <022> | Anna Brown | 30-39 | F | English | Welsh |
| <023> | Thomas Brown | 40-49 | M | English | London |

So the first two speaker turns of this transcription would be:

<022> okay so are we recording now?

<023> yep

Never leave out the initial zeroes – for our work, <022> and <22> are not the same thing!

When transcribing, please make sure that a new speaker starts on a new line, leaving one line between. Speaker tags should not appear in any position other than at the start of a new line.

If you think a particular utterance is said by a speaker but you aren't sure, please indicate this with e.g. <003?>. If you aren't sure who is speaking please identify the speaker as male or female by using <M> and <F> respectively. **Never use <M?> or <F?>.**

6. i) Multiple speakers

If multiple speakers say exactly the same thing at the same time, please write this as <MANY>. For example, if a whole class respond to a teacher's question with the answer "Friday", then this would be written:

<001> what day is the homework due in?

<MANY> Friday

7. ANONYMIZATION

7.1 People

Anonymize names of people. All anonymized names should include a gender tag (*male*, *female* or *neutral*). Indicate the gender of the name where possible, e.g.

"Dave" becomes <name M>

"Susan" becomes <name F>

If gender cannot be interpreted, either from the name (e.g. "Alex", which could be either "Alexandra" or "Alexander") or from the context, then use <name N>. This includes instances where just a family name, and no personal name, is given, or cases where a family name applies to a mixed-sex group, e.g. "Mr and Mrs Jones".

Where an anonymised name is more than one word long, only use a single <name ...> code (e.g. <name M> or <name F>).

| What is spoken | What is transcribed |
|---------------------------------------|-------------------------------------|
| my sister Briar Rose is older than me | my sister <name F> is older than me |
| goodbye Dr Wentwood-Smythe | goodbye Dr <name N>” |
| Jean-Pierre Duroy is horrible | <name M> is horrible |
| we invited Mr and Ms Smith | we invited Mr and Ms <name N> |

BUT:

“we invited Robert and Harriet Smith” → “we invited <name M> and <name F>”
(no extra <name N> for “Smith”).

7.2 Places

Anonymize names of locations and institutions/businesses which you judge to be locally identifiable, i.e. locations which are so specific that anyone reading the transcription could, with fairly little effort, use this information to help identify the speaker or someone who the speaker is talking about.

“I saw him at the Royal Tavern” becomes “I saw him at the <place>”

As in the example above, if the name of the place comprises several words (i.e. the “Royal Tavern”, which is the name of a pub and contains two words), do not attempt to retain such linguistic information (e.g. by transcribing “I saw him at the <place> <place>”). Simply the <place> tag, regardless of the length of the place name, is adequate. The exception to this is if the place is described as being located within another, separate, identifiable location – for example

“I saw him at the Royal Tavern in Blyth” becomes “I saw him at the <place> in <place>”

Note that this rule only applies to *names* of such locations and institutions, and not other associated words that on their own cannot identify the place. So, if somebody mentions that their child goes to a certain school, the level of anonymization depends on what is said:

“My daughter goes to Plessey road first school in Blyth” becomes “My daughter goes to <place> first school in <place>”

Likewise

“My daughter goes to first school in Blyth” becomes “My daughter goes to first school in <place>”, and NOT “My daughter goes to <place> in <place>”.

In no case does the word “school” need to be anonymised as it is not part of the identifiable place-name.

Do not anonymize the names of locations which are so general that they would not help a user of the corpus to identify any of the speakers etc. in the corpus.

“We went on holiday to France” – this would not be anonymized

7.3 Famous people

Do not anonymize the names of famous people, which are so general that they would not help a user of the corpus to identify any of the speakers etc. in the corpus.

“Did you see David Cameron’s speech last night?” – this would not be anonymized

7.4 Personal information

Anonymize personal information. Here is a full list of personal information anonymization tags:

| | |
|---|-----------|
| Telephone numbers (this includes all types – landline, mobile etc.) | <tel-num> |
| Addresses (any address which is spoken – this includes postcodes) | <address> |
| Email addresses | <email> |

| | |
|---|------------|
| Bank details (card numbers, account numbers, sort codes, etc.) | <bank-num> |
| Social media username (e.g. Twitter handles, skype names) | <soc-med> |
| Date of birth | <DOB> |
| Other personal information which is not captured by any of the above categories | <pers-inf> |

8. UTTERANCES

It can be very difficult to decide where to put sentence-breaks into spoken recordings, particularly when the speakers may talk for a long time with few pauses and numerous changes of topic. For this reason we think of the speech in *utterances*, rather than in 'sentences'. An utterance is the length of a speaker turn – that is, we do not break the speaker turn down further into sentences.

Utterances **should not start with a capital letter nor end with a full stop**:

<001> so I was thinking that erm it would be a good idea to decide on the procedure and then discuss it today

Another exception is where the speaker asks a question and then carries on talking. A question mark should be used, but no capital letter should be used at the beginning of the following utterance (unless it is a proper noun or / see: 14. ACRONYMS, SPELLING AND CAPITALISATION)

<001> what do you reckon? I think we should definitely go

9. PUNCTUATION

No punctuation should be used in transcription i.e. no commas, colons, dashes or full stops. Brackets – round, square and angled – are used to indicate tags and therefore should *never* be used in the 'normal' way.

Utterance ends **should not** be marked with a full stop.

Abbreviations and short forms should not be followed by a full stop: Dr Green not Dr. Green, Ms Black not Ms. Black.

Never use quotation marks of any kind.

There are two exceptions to this no-punctuation rule:

1. Question marks, which should be used for:

- (1) obvious questions (either *yes-no* questions or *who- what-where-when-how-why-type* questions: e.g. *are you happy? what did you do?*)
- (2) rhetorical questions
- (3) tag questions (e.g. *I told you didn't I?*)
- (4) statements with obvious rising intonation.

These are exemplified below:

| | |
|---|---|
| <001> sorry I didn't know you were moving it | = <i>Not a question</i> |
| <002> well what did you think I was trying to do? | = <i>Obvious question</i> |
| <001> you were bending down to have a look at it? | = <i>Statement with rising intonation (giving it question function)</i> |
| 002> do I look like an idiot? | = <i>Rhetorical question</i> |
| <001> you're not angry (.) are you? | = <i>Tag question</i> |

If a question utterance is interrupted or incomplete, only use a question mark at the end

<001> is this

<002> I don't know

<001> important?

2. Hyphens, which should be used for:

- (1) Proper nouns (e.g. Hay-on-Wye)
- (2) Numbers (e.g. one hundred and forty-six, four-year-old)
Only numbers between 21 and 99 should be hyphenated.

10. PAUSES

Do not record pauses that come at the beginning of an utterance. Record short and long pauses that occur during utterances; only recording pauses that occur between utterances if they are long. The long pauses between utterances should be recorded at the end of the first utterance in the pair.

| | | |
|-------------|-------|--|
| Short pause | (.) | Only use this tag for pauses which are between one second and five seconds, and only which occur during utterances. Do not record pauses which are less than one second. |
| Long pause | (...) | Use this tag for any pauses which are over five seconds, either during or between utterances. |

E.g.

<001> I had pizza and (.) chips last night

Short pause marked where it occurs during the utterance

E.g.

<001> I can't believe (...) I can't believe you just said that

Long pause marked where it occurs during the utterance

E.g.

<001> did you enjoy the film? (...)

<002> well erm not really actually

Long pause, which occurs between the two utterances, is marked at the end of the first utterance

11. UNFINISHED WORDS (FALSE STARTS)

A speaker may often begin a word but may not finish it. Please use the equals sign to mark where a word is unfinished:

<001> yes he's a ba=bachelor

<003> the test results were in=inc=inconclusive

| Feature | Transcription guideline | Example |
|--|--|--|
| False starts and repairs | Mark these using the equals sign (no space before or after) | Au=Au=August |
| Truncated words (not subsequently completed) | Mark these using the equals sign (no space before, space after) | it resem= (.) looks like (only include a pause here, if there's a gap between the truncated word and the next). |

12. OVERLAPS

Where one speaker interrupts another or tries to join in the conversation and the speech overlaps, use (this is a capital 'o', not a zero). This tag should be used **only at the start** of the turn of the speaker who is interrupting:

<001> erm a famous person whose name is Anne Hathaway

<002> okay can you tell us a bit about her?

<001> er er she she is a very famous movie star in America

Here when speaker one says “er er she she is a very famous movie star in America” this overlaps with speaker two saying “okay can you tell us a bit about her”. The exact position of the start of the overlap in the speech of speaker two does not need to be recorded. You do not need to mark the end of the overlap.

13. UNINTELLIGIBLE SPEECH / GUESSES

Where a speaker is unclear but you are able to have a guess at what the speaker is saying, use <u=GUESSEDWORDS> (where GUESSEDWORDS should be replaced by your guess!)

Please **do attempt** to guess the word or words you hear, and indicate them like this:

<001> this was relevant to the <u=manipulation> of the characters

If you can't make a guess as to what the word is, use <u=?>:

<001> this was relevant to the <u=?> of the characters

Please also try to make informed guesses. For example, in a transcription about farming you might hear the following utterance:

<001> yes, a lot of my work involves <u=?>

If you could make out the first letter of the utterance as 'm', some reasonable guesses may be *milking*, *mowing*, *mixing* depending on what had previously been said and what came next.

Please try and think about the context and about the topic – it may be useful to check back over the guesses you've made once you've got to the end of the recording – often the subject becomes clearer as the recording goes on.

It may also be useful to check something you aren't sure of (sometimes an unusual name or place) using Google – we don't expect you to spend lots of time doing this, but it can often be a quick solution and a good way to double check things.

14. ACRONYMS/SPELLING/CAPITALISATION

Names of people, places, companies, organisations, institutions, and book or publication titles, should always be capitalised and hyphenated in the usual way:

Mrs Jones, Steve Smith, Lancaster, Southend-on-Sea, The Catcher in the Rye, etc.

Use word-initial-capital for proper nouns and “I”. If a proper noun includes a number, it is spelled out (see 15. NUMBERS).

Proper nouns include:

| | |
|---|--|
| Names of people (but see <u>7. ANONYMIZATION</u>) | Roger, Shakespeare, Punch and Judy |
| Place names and derivatives (but not “little words” like “the, of, a”) | England, English, North Sea, Mars, Statue of Liberty, the London Eye |
| Names of products and institutions (the initial letter of each word is capitalised regardless of the official spelling, see <i>Iphone</i>) | Google, Facebook, Iphone, Microsoft, American Broadcasting Company, University of Vienna |
| Religions, religious institutions and derivatives | Christianity, Buddhism, Catholicism, Catholic, Buddhist |
| Names of days, months and festivals | Monday, February, Christmas, Chinese New Year, Hanukkah |

Not capitalised

- No capitalisation is used for titles or ‘honorific’ uses: archbishop, pope, king, duke, god, doctor, reverend, her majesty, his highness
- No capitalisation is used when originally proper nouns are employed as common nouns or verbs: I googled this, he was facebooking, she tweeted that she had no time

Only use abbreviations for the following titles: Mr, Ms, Mrs, Miss, Master, Dr

All other personal titles: write as normal words; do not abbreviate and do not capitalise.

Examples:

| | |
|------------------------------|---|
| Police / military: | superintendent <name F>, captain <name N> |
| Religious (historical): | guru Nanak, prophet Muhammad, saint John the baptist |
| Religious (contemporary): | reverend <name F>, ayatollah Khomeini, archbishop Rowan Williams |
| Professional: | professor <name F>, <name M> esquire, Dr <name F> BA MA PhD (for this last one see also acronym rules below). |
| Political: | chairman Mao, president Bush, lord justice Smythe |
| Aristocratic: | queen Elizabeth the second, king Ethelred the unready, duke Richard of York, lord and lady <name N>, sir Walter Raleigh, emperor Caligula, prince Albert. |

Hopefully these will be rare!

Please use established conventions for writing acronyms, but do not include dots:

<001> he's staying at the YMCA next week

<001> I've just bought it on DVD

<001> I'm going to the USA for twelve months

Plural forms should have a small 's' and no apostrophe:

<001> there were three PhDs awarded

<001> I've got so many CDs I don't know where to put them

Past tense forms should have an apostrophe and a small 'd':

<001> he MOT'd his car last week

When a speaker is clearly spelling something out in letters, (and not using an acronym), these should be written in capitals with a space between them:

<001> I said no that's N O

<001> my name is Bronwyn that's B R O N W Y N

Please do not put spaces between the letters of acronyms.

<001> we ourselves us that's spelt U S (.) us

– *spelt out word, space between U & S*

<001> I spent a month in the US

– *acronym, no space between U & S*

Finally, always write okay and not OK, O.K. or O K.

15. NUMBERS

Please write numbers out as words. If the speaker pronounces the number 0 as 'oh' then please write 0 (i.e. the number '0'). If the speaker pronounces 0 as 'zero' then please write 'zero'.

Dates should also be written how they're spoken. Numbers such as 31, 26, and 58 should be written in hyphenated form: one thousand and twenty-six, a hundred and two, zero, two double 0 five, twenty-first of the twelfth nineteen eighty-three

Times should be written out in words: twelve o'clock, five thirty, nine o'clock, half past two, ten to eleven

There are a few exceptions where numbers should be written as figures, including:

A4 paper

3D

MP3

Road names e.g. A66, A1

16. NONSTANDARD WORDS OR SOUNDS

Use the following spellings for nonstandard verbalisations (so-called ums and ers" or filled-pauses"):

| What it sounds like | How to write it |
|--|-----------------|
| Has the vowel found in "father" or a similar vowel; usually = realisation, frustration or pain | ah |
| Has the vowel found in "road" or a similar vowel; usually = mild surprise or upset | oh |
| Has the vowel in "bed" or the vowel in "made" or something similar, without an "R" or "M" sound at the end; usually = uncertainty, or 'please say again?' | eh |

| | |
|--|-----|
| A long or short “er” or “uh” vowel, as in “bird”; there may or may not be an “R” sound at the end; usually = uncertainty | er |
| As for “er” but ends as a nasal sound | erm |
| Has a nasally “M” or “N” sound from start to end; usually = agreement | mm |
| Like an “er” but with a clear “H” sound at the start; usually = surprise | huh |
| Two shortened “uh” or “er”-type vowels with an “H” sound between them, usually = disagreement; OR, a sound like the word “ahah!”; usually = success or realisation | uhu |

Please use **only** the spellings listed above.

If you hear a noise that does not match one of this list of 8 possible spellings, use the **closest-sounding** spelling from the list.

- For example, ‘mm’ should be used to cover all kinds of nasal-sounding agreement noises of various lengths, including but not limited to: *mm*, *mmm*, *mm-mm*, *mm-hm*, etc.
- Likewise, use ‘eh’ for sounds like *eee*, *ey*.
- Use ‘er’ also for *uh*, *ughh*
- Use ‘ah’ also for a pained *aaaargggghhhh*, for an *awwww!* ‘Isn’t-that-cute’ type of noise, or even for a pirate’s *Arrrrr!* (And so on.)

17. NONSTANDARD CONTRACTIONS OR SHORTENINGS

Please do not correct contractions that are acceptable in Standard English. E.g. don’t change contractions such as: *he’s*, *I’ve*, *we’re*, *I’m*, *don’t*, *she’ll* etc.

Use the following conventions and spellings for standard contractions: *ain’t*, *aren’t*, *can’t*, *cos*, *couldn’t*, *couldn’t’ve*, *daren’t*, *daren’t’ve*, *didn’t*, *doesn’t*, *don’t*, *hadn’t*, *hasn’t*, *haven’t*, *he’d*, *he’s*, *I’d*, *I’m*, *isn’t*, *it’s*, *I’ve*, *ma’am*, *may’ve*, *might’ve*, *mightn’t*, *mightn’t’ve*, *mustn’t*, *mustn’t’ve*, *must’ve*, *needn’t*, *needn’t’ve*, *oughtn’t*, *shan’t*, *she’d*, *she’s*, *shouldn’t*, *shouldn’t’ve*, , *wasn’t*, *weren’t*, *we’ve*, *won’t*, *wouldn’t*, *wouldn’t’ve*, *you’d*, *you’ve*

Plus, also ‘d, ‘s, ‘re, ‘ll, ‘ve, ‘d’ve, ‘ll’ve can attach to many words as standard contractions for very common words. Use the standard contraction spelling if you are confident that the pronunciation is as shortened as possible, down to just a very short vowel and consonant, or even less:

| | |
|--------|--|
| ‘d | <i>had</i> or <i>would</i> (see also below on “MOT’d”) |
| ‘s | <p><i>is, has</i> or <i>possessive</i>.</p> <p>Remember, however, the standard way of typing the possessive is ‘s for a singular word and s’ for a word ending in plural ‘s’. (E.g. The dog’s tail. The dogs’ basket. The fox’s nose. The foxes’ food-source. The church’s tower. The churches’ collaboration.)</p> <p>BUT <i>it+possessive</i> = <i>its</i> not <i>it’s</i>. (E.g. it’s funny that its head fell off)</p> <p>BUT <i>who+possessive</i> = <i>whose</i> not <i>who’s</i>. (E.g. who’s that? the man whose bike you stole)</p> |
| ‘re | <i>are</i> |
| ‘ve | <p><i>have</i></p> <p>Be very careful not to confuse “of” and “ve” which sound the same (just a very short “uhv”) when pronounced quickly). It should always be “would’ve” not “would of” for instance.</p> |
| ‘ll | <i>will</i> or <i>shall</i> |
| ‘d’ve | <i>would have</i> |
| ‘ll’ve | <i>will have</i> |

Examples: *The women’ll’ve done it, they’ll’ve left ages ago, I’d’ve been happy*

There are some semi-standard merged words: *dunno, gonna, wanna, gotta, kinda, sorta*

These should be used provided that it's very clear that speakers are saying, e.g. gonna, dunno or wanna rather than going to, don't know or want to. If you're unsure, please use the standard form.

18. SPEAKER ACCENT/DIALECT

Apart from the specific list above, do not make distinctions that are based only on how the speaker pronounces a word. For example if you hear a word with first or last consonant silent due to fast speech, don't leave out that letter. If you hear a vowel pronounced differently due to accent, don't write it differently:

- Don't use *hooose*: should be *house* even if it sounds like OO instead of OH
- Don't use *goiin* : should be *going* even with silent G
- Don't use *fish an chips*: should be *fish and chips* even with silent D
- Don't use *im, ospital, appy*: should be *him, hospital, happy* even with silent H
- Don't use *me* if the speaker is saying *my*, e.g. if the speaker says **have you seen me hat?* then you should write it as *my* not as *me*.
- Don't use *whatevva*: should be *whatever* even with an "Ah" sound at the end
- Don't use *somefink*: should always be *something* even with an "F" sound
- Don't use *dese/dose*: should always be *these/those* even with a "D" sound
- Don't use *bover*: should always be *bother* even with a "V" sound

The exception is that a Southern/London dialect might use "innit". Like the contractions above, only use *innit* if you are sure: otherwise use either *isn't it* or *ain't it*.

19. NON-LINGUISTIC VOCALISATIONS

Non-verbal vocalisations such as coughing, laughter etc. are marked with square brackets. Please use the following conventions.

| Category | Example | Comments |
|----------|---------|---|
| Laughter | [laugh] | When only one speaker laughs include this where the laugh occurs in their speaker turn. When more than one speaker laughs, give on a separate line. Only use <i>laugh</i> , i.e. don't use <i>giggle</i> , <i>chuckle</i> |

| | | |
|-----------------------------------|--|---|
| Coughs, clearing throat, gasps... | [cough] [gasp] [sneeze] [sigh] [yawn] [whistle] [misc] | Include in the speaker turn. Don't use a code for humming – use the “mm” introduced above. Don't use a code for screaming or yelling wordlessly – use the “ah” introduced above. Misc = any noise clearly produced by a human mouth that you can't easily describe |
| Singing | [sing=LYRICS] | The word LYRICS should be replaced by anything that is sung by the speaker, e.g.: <001> it's a song that goes [sing=somewhere over the rainbow way up high] |

| Other non-English sounds/speech | Example | comments |
|---------------------------------|--|---|
| Foreign languages | [f=French= ou est la gare?] [f=French] [f=?=kooda hafeez] [f=?] | <p>The format is: [f=LANGUAGE=WORDS], where LANGUAGE should be replaced with e.g. <i>French</i> or <i>Spanish</i> etc., and WORDS with the words that are spoken.</p> <p>If the LANGUAGE is unknown, use [f=?]</p> <p>If the WORDS can't be transcribed by you, leave out the =WORDS part.</p> <p>What we would expect you to transcribe:</p> <ul style="list-style-type: none"> - Some foreign words are commonly used by English speakers, so these can be transcribed without this tag e.g. “ah well c'est la vie”. - If it is easy for you to have a good guess at a representative spelling as you heard it e.g. “kooda hafeez” “in weeno werritass” <p>DO NOT do this unless it is easy! It is fine to just use [f=?]</p> |

| | | |
|--------------------------|------------|--|
| Nonsense / made-up words | [nonsense] | Only use this tag if the speaker is obviously not using a foreign language. If they are using made-up words which can be transcribed phonetically, then do this instead with no codes, e.g. <002> yes indeed. indeedilydoodily. |
|--------------------------|------------|--|

You should not make up new types of noise **unless it is absolutely necessary**.

Never include comments about how a sentence is said e.g. “enthusiastically” or “exaggerated”.

20. EVENTS

An “event” is anything audible **and relevant** on the recording that is not produced by voices of the speakers you are transcribing. The [e=SOMETHING] tag represents events, where SOMETHING is replaced by the type of event. Like the long pauses between utterances, events which occur between utterances are to be recorded at the end of the preceding utterance, rather than on a separate line.

E.g.

<001> I had a lovely time [e=sound of phone]

<002> oh I’ll go and get that

You do not have to code every single noise. The general rule is it must be a **relevant** event. The detailed rules for different events are given below.

| | | |
|-------------------|---------------------|--|
| Background speech | [e=background talk] | Use this when there is a general noise of conversation e.g. chatting before a lecture. |
|-------------------|---------------------|--|

| | | |
|-----------------------------|---|---|
| Unintelligible conversation | <p>[e=unintelligible]</p> <p><001> where did you go on your holiday?</p> <p>[e=unintelligible]</p> <p><002> oh yes sounds like you had a good time</p> | <p>Use this when the main participants in the recording are carrying out a conversation or conversations – lasting 2+ speaker turns - which cannot be heard clearly.</p> <p>Also use when all speakers talk together and individual speakers cannot be distinguished.</p> <p>(NB do not use this for individual speaker turns which cannot be heard, instead use <u> tags).</p> <p>Use this when you can't distinguish the speakers, therefore this tag is on a separate line with no speaker ID.</p> |
| Overlapping exchanges | <p>[e=begin overlap]</p> <p><001> so where do you think it will take place?</p> <p><003> in the lecture theatre probably</p> <p><001> I suppose so</p> <p>[e=end overlap]</p> | <p>In some files, groups of speakers hold different conversations at the same time. If both conversations are audible, please write the separate conversations out one at a time. This enables you to keep the corresponding speaker turns together, so that each conversation makes sense when you read it. Include the relevant [e=...] tag at the beginning and the end of the overlapping section.</p> |

| | | |
|--|--|--|
| Sounds and noises | <p>[e=sound of X]</p> <p>X can be ...</p> <ul style="list-style-type: none"> • car, i.e. [e=sound of car] • shouting i.e. [e=sound of shouting] • phone • applause • machinery • animal • siren <p><i>only add to this list if absolutely necessary, and try to use as few words as possible.</i></p> | <p>Only include sounds which affect or disrupt the conversation. Give in the format sound of ...”.</p> <p>If it occurs while someone is speaking, include in the speaker turn. If it occurs between speaker turns, give on a separate line.</p> <p>Never add extra detail to the description – just the bare statement of what it is.</p> |
| Music | [e=music] | <p>Only include music which affects or disrupts the conversation. Do not include type of music or song title.</p> |
| Abrupt end of recording | [e=abrupt end] | <p>Only use if a recording ends mid-word or mid-sentence. Type on a new line.</p> |
| People entering and leaving conversation venue | [e=001 leaves] | <p>Only include if the conversation is affected.</p> <p>Don’t mark people entering the room. Ignore movements of people other than your conversation participants.</p> |
| Problems in recording | [e=recording skips] | <p>Only use if a whole speaker turn (or more) is affected and the conversation no longer joins up correctly. If only a few words are unintelligible, use the <u...> tag.</p> |

21. STANDARD SPELLINGS

✓ etcetera

- ✓ alright
- ✓ okay
- ✓ whisky
- ✓ racket
- ✓ email
- ✓ realised – any words that can be written with an 's' or 'z' use the 's' form.
- ✓ Woah
- ✓ Grandad
- ✓ Summat (not summit, careful with global change on summit as a mountain)
- ✓ Couple of - not coupla
- ✓ Lot of not lotta
- ✓ Out of not outa
- ✓ No (not nah, na)

Appendix D. DTD for the Spoken BNC2014

The following DTD is also included in the XML download of the corpus, filename ***bnc2014spoken.dtd*** .

<!--

Document Type Definition for

```
=====
The SPOKEN BRITISH NATIONAL CORPUS 2014
=====
```

This DTD draws on several examples and best practices. The two primary sources are (a) the Text Encoding Initiative (TEI), and (b) the variation of TEI that is used by the original 1994 British National Corpus.

However, the practices codified in this DTD differ from those in TEI in a number of ways. Most notably, the XML markup of the BNC2014 is dramatically simplified relative to TEI/BNC1994. This is in line with the recommendations of Hardie (2014), who argues that a small subset of the features of XML and TEI are more than sufficient for most corpus linguistic analysis.

Reference:

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

For that reason, while much of what follows will be familiar to anyone who has used TEI, it does not actually use a TEI definition. Nor are any of the advanced features of XML/DTD used.

CHANGELOG

=====

2018-09-01 Added header elements (AH)
2016-01-25 Adjusted possible locations for event/vocal tags (AH, RL)
2015-02-03 Initial version created (AH)

-->

<!--

The text element - root for each text in the corpus.

-->

<!ELEMENT text (header,body)>

<!ATTLIST text

id

ID

#REQUIRED

```

>

<!--
This header definition covers the speaker metadata, which is included within it.
Element names are the column codes used in the CQP{web installation of the corpus.
-->

<!ELEMENT header
      (rec_length,rec_date,rec_year,rec_period,n_speakers,list_speakers,rec_loc,relationships,
       topics,activity,conv_type,conventions,in_sample,transcriber,speakerInfo) >

<!ELEMENT rec_length      (#PCDATA)>
<!ELEMENT rec_date        (#PCDATA)>
<!ELEMENT rec_year        (#PCDATA)>
<!ELEMENT rec_period      (#PCDATA)>
<!ELEMENT n_speakers      (#PCDATA)>
<!ELEMENT list_speakers   (#PCDATA)>
<!ELEMENT rec_loc         (#PCDATA)>
<!ELEMENT relationships   (#PCDATA)>
<!ELEMENT topics          (#PCDATA)>
<!ELEMENT activity        (#PCDATA)>
<!ELEMENT conv_type       (#PCDATA)>
<!ELEMENT conventions     (#PCDATA)>
<!ELEMENT in_sample       (#PCDATA)>
<!ELEMENT transcriber     (#PCDATA)>

<!ELEMENT speakerInfo     (speaker)*>

<!ELEMENT speaker
      (exactage,age1994,agerange,gender,nat,birthplace,birthcountry,l1,lingorig,diaclect_rep,ha
       b_city,hab_country,hab_dur,diaclect_l1,diaclect_l2,diaclect_l3,diaclect_l4,edqual,occupation
       ,socgrade,nssec,l2,fls,in_core) >

<!ATTLIST speaker
      id          ID          #REQUIRED
>

<!ELEMENT exactage      (#PCDATA)>
<!ELEMENT age1994       (#PCDATA)>
<!ELEMENT agerange      (#PCDATA)>
<!ELEMENT gender        (#PCDATA)>
<!ELEMENT nat           (#PCDATA)>
<!ELEMENT birthplace    (#PCDATA)>
<!ELEMENT birthcountry  (#PCDATA)>
<!ELEMENT l1            (#PCDATA)>
<!ELEMENT lingorig      (#PCDATA)>
<!ELEMENT diaclect_rep  (#PCDATA)>
<!ELEMENT hab_city      (#PCDATA)>
<!ELEMENT hab_country   (#PCDATA)>
<!ELEMENT hab_dur       (#PCDATA)>
<!ELEMENT diaclect_l1   (#PCDATA)>
<!ELEMENT diaclect_l2   (#PCDATA)>
<!ELEMENT diaclect_l3   (#PCDATA)>
<!ELEMENT diaclect_l4   (#PCDATA)>

```

```

<!ELEMENT edqual      (#PCDATA)>
<!ELEMENT occupation  (#PCDATA)>
<!ELEMENT socgrade    (#PCDATA)>
<!ELEMENT nssec       (#PCDATA)>
<!ELEMENT l2          (#PCDATA)>
<!ELEMENT fls         (#PCDATA)>
<!ELEMENT in_core     (#PCDATA)>

<!--
The body contains everything in the actual text itself.
This is a string of utterances, potentially with vocalisations/events/pauses
that can't be linked to a particular utterance inbetween.
-->
<!ELEMENT body      (u|event|pause)*>

<!--
u is the main data-containing element, thus its complex contents and attribute list.
-->
<!ELEMENT u      (#PCDATA|vocal|event|anon|pause|shift|unclear|trunc|foreign)*>
<!--
    note 1: ultimately "who" will be an IDREF, but until we have speaker IDs in the header,
    we define it as NMTOKEN.
    note 2: we have borrowed "overlap/smooth" from TEI; TEI also allows "latching/pause",
    but we never use these.
-->
<!ATTLIST u
    who          NMTOKEN          #REQUIRED
    whoConfidence (high|low)      "high"
    trans        (overlap|smooth) "smooth"
    n            CDATA            #IMPLIED
>

<!ELEMENT vocal EMPTY>
<!ATTLIST vocal
    desc          (laugh|cough|gasp|sneeze|sigh|yawn|whistle|nonsense|misc)
                  #REQUIRED
>

<!--
Unlike vocals, events can have anything in their description.
-->
<!ELEMENT event EMPTY>
<!ATTLIST event
    desc          CDATA          #REQUIRED
>

<!ELEMENT anon EMPTY>
<!ATTLIST anon
    type
    (name|place|telephoneNumber|address|email|financialDetails|socialMediaName|dateOfBirth|m
    iscPersonalInfo)

```

```

                                #REQUIRED
nameType          (m|f|n)      #IMPLIED
>
<!-- NB m = name identifiable as that of a male, f = ditto for a female, n = name whose
      owner's gender is not determinable -->

<!ELEMENT pause EMPTY>
<!ATTLIST pause
      dur          (long|short)      #REQUIRED
>
<!-- NB for our purposes, short = up to 5 seconds; long = more than that -->

<!--
shift has been borrowed from the BNC1994 - it is similar to, but not the same as,
the shift element in TEI. However, the BNC1994 uses it for all kinds of voice quality
shifts, whereas we only use it to indicate shifts to sung-lyrics and back again.
-->
<!ELEMENT shift EMPTY>
<!ATTLIST shift
      new          (singing|normal)  #REQUIRED
>

<!--
An unclear can either be a point, or it may contain PCDATA.
It can therefore also contain a subset of other elements from <u>.
-->
<!ELEMENT unclear (#PCDATA|vocal|anon|pause|trunc)*>

<!--
Truncated word, i.e. incomplete, false starts, repairs, etc.
-->
<!ELEMENT trunc (#PCDATA)>

<!--
Text identified as non-English.
Can contain only a subset of the elements from <u>.
Especially note: unclear is allowed in foreign, but foreign is not allowed in unclear.
-->
<!ELEMENT foreign (#PCDATA|vocal|anon|pause|unclear|trunc)*>
<!-- any foreign must have a lang. In theory, we could enumerate these
      (as it must be an ISO 639-2 code, including "und" for undetermined if unknown)
      but that would take up too much space... so, we just leave it as a CDATA.
-->
<!ATTLIST foreign
      lang          CDATA            #REQUIRED
>

<!-- End of DTD -->

```

Appendix E. DTD for the Written BNC2014

This will be added upon the release of the XML version of the Written BNC2014.

Appendix F. UCREL CLAWS6 tagset (<http://ucrel.lancs.ac.uk/claws6tags.html>)

| | |
|-------|--|
| APPG | possessive pronoun, pre-nominal (e.g. my, your, our) |
| AT | article (e.g. the, no) |
| ATI | singular article (e.g. a, an, every) |
| BCL | before-clause marker (e.g. in order (that), in order (to)) |
| CC | coordinating conjunction (e.g. and, or) |
| CCB | adversative coordinating conjunction (but) |
| CS | subordinating conjunction (e.g. if, because, unless, so, for) |
| CSA | as (as conjunction) |
| CSN | than (as conjunction) |
| CST | that (as conjunction) |
| CSW | whether (as conjunction) |
| DA | after-determiner or post-determiner capable of pronominal function (e.g. such, former, same) |
| DA1 | singular after-determiner (e.g. little, much) |
| DA2 | plural after-determiner (e.g. few, several, many) |
| DAR | comparative after-determiner (e.g. more, less, fewer) |
| DAT | superlative after-determiner (e.g. most, least, fewest) |
| DB | before determiner or pre-determiner capable of pronominal function (all, half) |
| DB2 | plural before-determiner (both) |
| DD | determiner (capable of pronominal function) (e.g. any, some) |
| DD1 | singular determiner (e.g. this, that, another) |
| DD2 | plural determiner (these, those) |
| DDQ | wh-determiner (which, what) |
| DDQGE | wh-determiner, genitive (whose) |
| DDQV | wh-ever determiner, (whichever, whatever) |
| EX | existential there |
| FO | formula |
| FU | unclassified word |
| FW | foreign word |
| GE | germanic genitive marker - (' or's) |
| IF | for (as preposition) |
| II | general preposition |
| IO | of (as preposition) |
| IW | with, without (as prepositions) |
| JJ | general adjective |
| JJR | general comparative adjective (e.g. older, better, stronger) |
| JJT | general superlative adjective (e.g. oldest, best, strongest) |
| JK | catenative adjective (able in be able to, willing in be willing to) |
| MC | cardinal number, neutral for number (two, three..) |

| | |
|-------|---|
| MC1 | singular cardinal number (one) |
| MC2 | plural cardinal number (e.g. sixes, sevens) |
| MCGE | genitive cardinal number, neutral for number (two's, 100's) |
| MCMC | hyphenated number (40-50, 1770-1827) |
| MD | ordinal number (e.g. first, second, next, last) |
| MF | fraction, neutral for number (e.g. quarters, two-thirds) |
| ND1 | singular noun of direction (e.g. north, southeast) |
| NN | common noun, neutral for number (e.g. sheep, cod, headquarters) |
| NN1 | singular common noun (e.g. book, girl) |
| NN2 | plural common noun (e.g. books, girls) |
| NNA | following noun of title (e.g. M.A.) |
| NNB | preceding noun of title (e.g. Mr., Prof.) |
| NNL1 | singular locative noun (e.g. Island, Street) |
| NNL2 | plural locative noun (e.g. Islands, Streets) |
| NNO | numeral noun, neutral for number (e.g. dozen, hundred) |
| NNO2 | numeral noun, plural (e.g. hundreds, thousands) |
| NNT1 | temporal noun, singular (e.g. day, week, year) |
| NNT2 | temporal noun, plural (e.g. days, weeks, years) |
| NNU | unit of measurement, neutral for number (e.g. in, cc) |
| NNU1 | singular unit of measurement (e.g. inch, centimetre) |
| NNU2 | plural unit of measurement (e.g. ins., feet) |
| NP | proper noun, neutral for number (e.g. IBM, Andes) |
| NP1 | singular proper noun (e.g. London, Jane, Frederick) |
| NP2 | plural proper noun (e.g. Browns, Reagans, Koreas) |
| NPD1 | singular weekday noun (e.g. Sunday) |
| NPD2 | plural weekday noun (e.g. Sundays) |
| NPM1 | singular month noun (e.g. October) |
| NPM2 | plural month noun (e.g. Octobers) |
| PN | indefinite pronoun, neutral for number (none) |
| PN1 | indefinite pronoun, singular (e.g. anyone, everything, nobody, one) |
| PNQO | objective wh-pronoun (whom) |
| PNQS | subjective wh-pronoun (who) |
| PNQV | wh-ever pronoun (whoever) |
| PNX1 | reflexive indefinite pronoun (oneself) |
| PPGE | nominal possessive personal pronoun (e.g. mine, yours) |
| PPH1 | 3rd person sing. neuter personal pronoun (it) |
| PPHO1 | 3rd person sing. objective personal pronoun (him, her) |
| PPHO2 | 3rd person plural objective personal pronoun (them) |
| PPHS1 | 3rd person sing. subjective personal pronoun (he, she) |
| PPHS2 | 3rd person plural subjective personal pronoun (they) |

| | |
|-------|---|
| PPIO1 | 1st person sing. objective personal pronoun (me) |
| PPIO2 | 1st person plural objective personal pronoun (us) |
| PPIS1 | 1st person sing. subjective personal pronoun (I) |
| PPIS2 | 1st person plural subjective personal pronoun (we) |
| PPX1 | singular reflexive personal pronoun (e.g. yourself, itself) |
| PPX2 | plural reflexive personal pronoun (e.g. yourselves, themselves) |
| PPY | 2nd person personal pronoun (you) |
| RA | adverb, after nominal head (e.g. else, galore) |
| REX | adverb introducing appositional constructions (namely, e.g.) |
| RG | degree adverb (very, so, too) |
| RGQ | wh- degree adverb (how) |
| RGQV | wh-ever degree adverb (however) |
| RGR | comparative degree adverb (more, less) |
| RGT | superlative degree adverb (most, least) |
| RL | locative adverb (e.g. alongside, forward) |
| RP | prep. adverb, particle (e.g. about, in) |
| RPK | prep. adv., catenative (about in be about to) |
| RR | general adverb |
| RRQ | wh- general adverb (where, when, why, how) |
| RRQV | wh-ever general adverb (wherever, whenever) |
| RRR | comparative general adverb (e.g. better, longer) |
| RRT | superlative general adverb (e.g. best, longest) |
| RT | quasi-nominal adverb of time (e.g. now, tomorrow) |
| TO | infinitive marker (to) |
| UH | interjection (e.g. oh, yes, um) |
| VB0 | be, base form (finite i.e. imperative, subjunctive) |
| VBDR | were |
| VBDZ | was |
| VBG | being |
| VBI | be, infinitive (To be or not... It will be ..) |
| VBM | am |
| VBN | been |
| VBR | are |
| VBZ | is |
| VD0 | do, base form (finite) |
| VDD | did |
| VDG | doing |
| VDI | do, infinitive (I may do... To do...) |
| VDN | done |
| VDZ | does |

| | |
|-------|--|
| VH0 | have, base form (finite) |
| VHD | had (past tense) |
| VHG | having |
| VHI | have, infinitive |
| VHN | had (past participle) |
| VHZ | has |
| VM | modal auxiliary (can, will, would, etc.) |
| VMK | modal catenative (ought, used) |
| VV0 | base form of lexical verb (e.g. give, work) |
| VVD | past tense of lexical verb (e.g. gave, worked) |
| VVG | -ing participle of lexical verb (e.g. giving, working) |
| VVGK | -ing participle catenative (going in be going to) |
| VVI | infinitive (e.g. to give... It will work...) |
| VVN | past participle of lexical verb (e.g. given, worked) |
| VVNK | past participle catenative (e.g. bound in be bound to) |
| VVZ | -s form of lexical verb (e.g. gives, works) |
| XX | not, n't |
| YEX | punctuation tag - exclamation mark |
| YQUO | punctuation tag - quotes |
| YBL | punctuation tag - left bracket |
| YBR | punctuation tag - right bracket |
| YCOM | punctuation tag - comma |
| YDSH | punctuation tag - dash |
| YSTP | punctuation tag - full-stop |
| YLIP | punctuation tag - ellipsis |
| YCOL | punctuation tag - colon |
| YSCOL | punctuation tag - semicolon |
| YQUE | punctuation tag - question mark |
| ZZI | singular letter of the alphabet (e.g. A,b) |
| ZZZ | plural letter of the alphabet (e.g. A's, b's) |

Appendix G. The Oxford Simplified Tagset (adopted from <http://www.natcorp.ox.ac.uk/docs/URG/codes.html#klettpos>)

This table lists, for each of the twelve simplified wordclass tags used in the BNC1994 and subsequently adopted for the BNC2014, the corresponding CLAWS POS tags of which the class consists.

| Simple tag | significance | CLAWS C5 tags covered | CLAWS C6 tags covered |
|------------|--|---|--|
| ADJ | adjective | AJ0, AJC, AJS, CRD, DT0, ORD | All beginning in J or M, plus all beginning in D except DDQ, DDQGE, DDQV |
| ADV | adverb | AV0, AVP, AVQ, XX0 | All beginning with R, plus XX |
| ART | article | AT0 | All beginning in A except APPGE |
| CONJ | conjunction | CJC, CJS, CJT | All beginning in B or C |
| INTERJ | interjection | ITJ | UH |
| PREP | preposition | PRF, PRP, TO0 | All beginning in I, plus TO |
| PRON | pronoun | DPS, DTQ, EX0, PNI, PNP, PNQ, PNX | All beginning in P, plus APPGE, DDQ, DDQGE, DDQV, EX |
| STOP | punctuation | POS ¹⁵ , PUL, PUN, PUQ, PUR | All beginning in Y |
| SUBST | substantive (i.e. noun) | NN0, NNI, NN2, NP0, ONE ¹⁶ , ZZ0, NNI-NP0, NP0-NNI | All beginning in N or Z |
| UNC | unclassified, uncertain, or non-lexical word | UNC, AJ0-AV0, AV0-AJ0, AJ0-NNI, NNI-AJ0, AJ0-VVD, VVD-AJ0, AJ0-VVG, VVG-AJ0, AJ0-VVN, VVN-AJ0, AVP-PRP, PRP-AVP, AVQ-CJS, CJS-AVQ, CJS-PRP, PRP-CJS, CJT-DT0, DT0-CJT, CRD-PNI, PNI-CRD, NNI-VVB, VVB-NNI, NNI-VVG, VVG-NNI, NN2-VVZ, VVZ-NN2 | All beginning in F plus GE ¹⁷ |
| VERB | verb | VBB, VBD, VBG, VBI, VBN, VBZ, VDB, VDD, VDG, VDI, VDN, VDZ, VHB, VHD, VHG, VHI, VHN, VHZ, VM0, VVB, VVD, VVG, VVI, VVN, VVZ, VVD-VVN, VVN-VVD | All beginning in V |

¹⁵ While the BNC1994 documentation asserts that POS is grouped under STOP, in practice, it has been grouped under UNC in the XML Edition release of the BNC1994.

¹⁶ This tag only existed in pre-final versions of the C5 tagset.

¹⁷ For reasons of backward compatibility with the BNC1994 treatment of the equivalent tag POS.