Corpus manual of

The London-Lund Corpus 2

of spoken British English

Updated on 16 May 2019

Developers:

Nele Põldvere

Victoria Johansson

Carita Paradis

Contact information:

llc2@englund.lu.se

www.sol.lu.se/en/subjects/engelska/llc2

Contents

The manual: updates and changes

[To be written.]

Acknowledgements

A project of this size and nature would not have been possible without generous financial support from a number of foundations in Sweden and moral support from people who helped us in various stages of the project.

First and foremost, we are thankful to the Linnaeus Centre for Thinking in Time: Cognition, Communication, and Learning, financed by the Swedish Research Council (grant no. 349-2007-8695), and the Erik Philip-Sörensens Stiftelse, who financed the most challenging part of compiling LLC-2, namely the transcription of the recordings. The grants allowed us to hire a part-time research assistant, Paschal O'Hare, for 17 months in 2016–2017 and Nele Põldvere for seven months in 2018. The annotation of the transcriptions was carried out in 2019 and was financed by the Åke Wibergs Stiftelse.

Various foundations at the Centre for Languages and Literature at Lund University supported Nele Põldvere's data collection trips to the UK during 2015–2018. The five trips to the University College London and Lancaster University were supported by the Stiftelsen Olof Sagers Stipendiefond and the Birgit Rausing Language Programme, and the equipment and software used in the UK were purchased with the help of the department.

LLC-2 would not have become materialised without all the speakers who sacrificed their valuable free time and agreed to have their most private conversations recorded for language research and, in most cases, without any immediate gain or reward. Some of them did it out of empathy for fellow academics ("I know how hard it is to collect data"), but we can honestly say that all the speakers we recorded were genuinely curious about the project and invested in its success ("Let me know when the corpus is ready"). Their enthusiasm confirmed to us that corpus research is interesting, important and worth pursuing. Hopefully the speakers gained something from the project if only the chance to converse with the people around them as suggested by this extract from a conversation among two friends after 30 minutes of talking: "We should chat more often without like watching the TV instead". We are also grateful to the institutions, organisations and public figures who replied to our copyright requests and granted permission to use their material in the corpus, again without any monetary reward. Equally important to the project are the people who spread the word and helped us recruit speakers along the way.

Special thanks are due to Paschal O'Hare who undertook the strenuous work of transcribing spontaneous conversation and whose dedication and hard work resulted in the completion of almost half of the transcriptions in the corpus. We were glad to hear that, as a geologist by training, the most difficult word he came across in the recordings was a word

related to linguistics: morphophonology. Thanks are also due to Renata Kochančikaitė who helped with the transcriptions in December 2017.

From the start, LLC-2 was compiled with the intention of making it publicly available to the research community. Therefore, an important role is played by those who facilitated the release of the corpus for both offline and online use. First, we are grateful to the Lund University Humanities Lab for offering secure and long-term storage of LLC-2 on their corpus server. Thanks to Jens Larsson and Johan Frid, the corpus now has a home there, and thanks to Susanna Björverud, the corpus also has a homepage. Johan Frid also helped with anonymising the LLC-2 sound files so that they can be published alongside the transcriptions. Second, we are thankful to the CLARINO Centre Bergen and the University of Bergen for making LLC-2 available from *Corpuscle*. [Description of *Corpuscle.*] The launch of the corpus was celebrated in September 2019 with a one-day symposium at the Centre of Languages and Literature, called *Spoken language across time: celebrating the launch of the London-Lund Corpus 2*, which was jointly funded by the Birgit Rausing Language Programme and the Royal Swedish Academy of Sciences. We thank all the presenters and the audience for participating.

We are grateful to the staff, both past and present, at the Survey of English Usage at University College London for generously hosting Nele Põldvere during the four trips to London and for providing much-needed assistance in carrying out the recordings. They are: Bas Aarts, Sean Wallis, Kathryn Allan, Karen Dwyer, Ellen Smith-Dennis, Ian Cushing and, last but not least, Rachele De Felice. Rachele De Felice's contribution to the project cannot be overestimated; she worked on many fronts, from recruiting speakers to offering moral support, and we are extremely grateful for her unwavering support in the project. In the same way, we would like to extend our thanks to the ESRC Centre for Corpus Approaches to Social Science at Lancaster University for hosting Nele Põldvere during the summer of 2017 and Andrew Hardie for his assistance in marking up the corpus. We also appreciate the good advice given by Stefan Diemer from Saarland University on the best practices of recording Skype conversations. Last but not least, thanks to various foundations in Sweden, we have been able to participate in a number of conferences, workshops and symposia in Sweden and elsewhere and present the corpus and the research based on it to an international audience. We are thankful to the audience for their feedback, particularly in the 38th conference of ICAME in Prague in May 2017 where LLC-2 was first presented.

1. Preface

The London-Lund Corpus 2 (henceforth LLC-2) of spoken British English is a corpus of spoken language. It was developed at the Centre for Languages and Literature at Lund University, Sweden by Nele Põldvere, Victoria Johansson and Carita Paradis. Paschal O'Hare was employed as a research assistant.

The total number of orthographic words in LLC-2 is [number], which corresponds to [number] hours of recording. The corpus consists of [number] transcription files, each around 5,000 words in size, and [number] corresponding sound files. The corpus data were recorded during the period 2014–2019 in the UK (primarily in London) and in Lund, Sweden with [number] adult educated native speakers of British English. The corpus design includes seven broad text categories: face-to-face conversation, mobile phone/Skype conversation, broadcast discussions and interviews, parliamentary language, spontaneous commentary, legal language and prepared speech. For more information about the design of LLC-2, see Section 3 below.

LLC-2 was designed according to the same principles as the original London-Lund Corpus (henceforth LLC-1). LLC-1 was developed in collaboration of the Survey of English Usage at University College London and the Survey of Spoken English at Lund University. It was the world's first machine-readable corpus of spoken language with data recorded in the 1950s–1980s. The size and design of LLC-1 were taken as the ultimate goal in the compilation of LLC-2, which means that both corpora contain the same text categories and roughly to the same extent. The texts in LLC-2 were also designed to reflect the rapid technological advances of the late 20th and early 21st centuries, which explains the inclusion of communication and media channels not available at the time of compiling LLC-1. Section 4 provides further information about LLC-1 and the extent to which it is comparable to LLC-2.

The potential applications of LLC-2 are many. On the one hand, it allows researchers to study contemporary speech from a synchronic perspective and across different registers and groups of speakers. On the other hand, it facilitates principled comparisons across two different time periods of contemporary English with roughly 50 years in between. We hope that making LLC-2 accessible to the research community will stimulate research on spoken English in linguistics and related fields, both from a synchronic and a diachronic perspective.

LLC-2 can be accessed from two locations. First, since [date] the transcription files and the corresponding sound files are available for download from the Lund University Humanities Lab's corpus server at https://corpora.humlab.lu.se. The transcriptions are in XML (*eXtensible Markup Language*) format and time-aligned with the sound files. The

corpus is publicly available, but a password is needed to access the data. The password can be obtained by sending an email to llc2@englund.lu.se. Second, in early 2020, LLC-2 will also be accessible from the corpus management and analysis system *Corpuscle* at clarino.uib.no/korpuskel, developed and maintained at the CLARINO Centre Bergen in collaboration with the University of Bergen, Norway. *Corpuscle* enables the implementation of basic corpus linguistic techniques on LLC-2, from simple query searches to more complex concordance and collocation analyses. Both sources come fully equipped with metadata about the corpus texts and the speakers. Detailed instructions on how to access the corpus can be found in Section 2.

This corpus manual was designed to provide detailed information about the various stages of compiling LLC-2, from collecting data to eventually making the data available for public use. The size of the manual reflects the countless decisions and considerations that went into compiling the corpus, and the reader is encouraged to closely read through the manual to gain a solid understanding of the process. Alternatively, the reader is referred to the Quick Guide to Using LLC-2 in Section 9 below for an abridged version of the manual.

2. Access to the corpus

This section describes how to access LLC-2 from the Lund University Humanities Lab's corpus server (Section 2.1) and the online interface *Corpuscle* (Section 2.2). Furthermore, it presents the standard for acknowledging LLC-2 in any publication that arises from the use of the corpus (Section 2.3).

2.1 Download from the corpus server

LLC-2 is available for download from the Lund University Humanities Lab's corpus server at https://corpora.humlab.lu.se. The corpus server is based on software developed within The Language Archive (TLA) at Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands. All the corpora on the server, including LLC-2, are stored in accordance with the guidelines set by SWE-CLARIN, a Swedish member of the European Research Infrastructure for Language Resources and Technology.

The following files are available for download from the corpus server:

- [number] POS-tagged XML files,
- [number] XML files without POS-tagging,

- [number] corresponding sound files (in WAV or MP3 format),
- two Excel spreadsheets with metadata about the texts and the speakers (Read-only),
- a corpus manual,
- an End User Licence agreement.

All the transcription files in the corpus begin with 'T' followed by a unique numeric code (e.g., T001). The corresponding sound files begin with 'R'. For example, the transcription file T001 is connected to the sound file R001. The two Excel spreadsheets summarise the metadata information provided in the headers of each transcription file (see Section 8.1 below).

While the corpus manual and the End User Licence agreement are publicly visible, the rest of the corpus files are password-protected. To obtain the password, please read carefully the terms and conditions specified in the End User Licence agreement (see Appendix A), fill in the agreement (electronically or handwritten) and send it to llc2@englund.lu.se. If filled in by hand, the agreement should be scanned and attached to the email in PDF format. Please note that the password will be sent at the developers' earliest convenience.

2.2 The online interface Corpuscle

This section will be written when LLC-2 has been made publicly available on *Corpuscle* in early 2020.

2.3 How to acknowledge the use of the corpus

The corpus user is required to acknowledge LLC-2 in any publication or presentation arising from the use of the corpus data by adding the following reference (or a variation of it) to the reference list:

- Põldvere, N., Paradis, C., & Johansson, V. (2019). The London-Lund Corpus 2 of spoken British English. Lund, Sweden: Lund University. Available from https://corpora.humlab.lu.se

3. Corpus design

LLC-2 is a snapshot of spoken British English in 2014–2019. It covers a range of speech settings in which people participate, both as speakers and listeners, and the proportions of the

speech settings in the corpus are designed to reflect the proportions that exist in the population. For example, face-to-face conversation is given precedence over the other speech settings due to its predominance in the language used outside of the corpus. At the same time, LLC-2 is a comparable corpus and its design criteria follow closely those of LLC-1 and the decisions made some 50 years ago. With the aim of capturing the grammatical and stylistic variation that exists in English, LLC-1 was compiled to represent a number of different speech settings but without strict consideration of their statistical distribution in the population. For example, according to the developers of LLC-1, over 99% of speech is produced in face-to-face conversations among speakers who are on equal footing (Svartvik & Quirk, 1980, p. 11). This means that, in theory, the same proportion of the speech setting should exist in LLC-1. However, this is not the case. Face-to-face conversation makes up the largest proportion of the corpus but not all of it. Instead, room is left for other speech settings such as broadcast discussions and interviews, parliamentary debates and spontaneous commentary, just to name a few. Although only a select few of us have the authority and the expertise to host broadcast shows, debate in the Houses of Parliament or comment on a football match on live TV, many more people are exposed to these speech settings as listeners.

LLC-2 contains [number] orthographic words stored in [number] texts of around 5,000 words each. One text in the corpus is equivalent to either one single recording or multiple shorter recordings revolving around a similar subject matter and/or involving the same (one) speaker. Where possible, the recordings were transcribed in full; however, most of the texts in the corpus represent an excerpt from a recording. Considering that no part of a recording can be considered to be representative of the whole and that certain linguistic choices only become relevant in the beginning, middle or end of the speech setting, care was taken to retrieve the extracts from different parts of the recordings. Detailed information about each text is given in the metadata.

Table 1 below presents the complete design of LLC-2. As can be seen in the table, the corpus contains seven broad text categories: face-to-face conversation, mobile phone/Skype conversation, broadcast discussions and interviews, parliamentary language, spontaneous commentary, legal language and prepared speech. Most of the text categories are further divided into subcategories. For example, face-to-face conversations are divided into conversations among equals and disparates. The speakers are equal if they are friends, peers in the workplace or related by descent or marriage (e.g., parent–child, husband–wife), and they are disparates if they have hierarchically unequal positions in the workplace or the

educational institution (e.g., employer–employee in work meetings, teacher–student in tutorials). Mobile phone/Skype conversations in LLC-2 always take place among equals.[1] They are carried out either over a mobile phone or a computer; in the former, the speakers can only hear each other, but in the latter, they can also see each other through a webcam. In both cases, the speakers are not physically present in the same location. The rest of the text categories commonly involve public figures. Broadcast discussions and interviews are broadcast on TV, radio or the internet and include discussions and interviews on a specific topic. Parliamentary language is represented by question time where members of the House of Commons and the House of Lords ask questions from government ministers, and debates. Spontaneous commentary includes commentary on sports events and video games, and science and cooking demonstrations. Legal language is represented by hearings in the UK Supreme Court. Finally, prepared speech involves a single speaker and is given in a variety of settings. These include: political speeches, university lectures, popular science lectures and sermons. No attempts were made to distinguish between scripted and unscripted speech because of the blurred line between the two.

[1] In LLC-1, the distinction between conversations among equals and disparates is not limited

Table 1. Complete design of LLC-2 and the distribution of the text categories and subcategories in terms of the number of texts, words and speakers.

Text category	**Subcategory**	**Text IDs**	**#T**	**#T**	**#W**	**#W**	**#S**	**#S**
Face-to-face conversation	Equals	[…]	[…]	[…]	[…]	[…]	[…]	[…]
	Disparates	[…]	[…]		[…]		[…]	
Mobile phone/Skype conversation	Audio	[…]	[…]	[…]	[…]	[…]	[…]	[…]
	Video	[…]	[…]		[…]		[…]	
Broadcast discussions and interviews	Discussions	[…]	[…]	[…]	[…]	[…]	[…]	[…]
	Interviews	[…]	[…]		[…]		[…]	
Parliamentary language	Question time	[…]	[…]	[…]	[…]	[…]	[…]	[…]
	Debates	[…]	[…]		[…]		[…]	
Spontaneous commentary	Sports	[…]	[…]	[…]	[…]	[…]	[…]	[…]
	Video games	[…]	[…]		[…]		[…]	
	Science	[…]	[…]		[…]		[…]	
	Cooking	[…]	[…]		[…]		[…]	
Legal language	Hearings	[…]	[…]	[…]	[…]	[…]	[…]	[…]
Prepared speech	Politics	[…]	[…]	[…]	[…]	[…]	[…]	[…]
	Lectures	[…]	[…]		[…]		[…]	
	Popular science	[…]	[…]		[…]		[…]	
	Sermons	[…]	[…]		[…]		[…]	
			[…]	[…]	[…]	[…]	[…]	[…]

Note: '#T' refers to the number of texts, '#W' to the number of words and '#S' to the number of speakers. The first column of each denotation corresponds to the subcategory total and the second column to the text category total.

The corpus texts that fall under the seven text categories and their subcategories in LLC-2 are given in column 3 of Table 1. Furthermore, the table presents the distribution of the (sub)categories in the corpus in terms of the number of texts, words and speakers. [Description of the table.]

The seven text categories in Table 1 are characterised by a range of features that give rise to register variation and affect the way grammatical and lexical phenomena are distributed across varieties of language use. A distinction that is commonly made in corpus compilation is one between dialogue and monologue. The LLC-1 corpus manual, for example, reports that dialogues in LLC-1 are conversations and public discussions while monologues may be spontaneous or prepared (Greenbaum & Svartvik, 1990; see Section 4.1 below). According to the manual, then, the first three text categories in LLC-2 as reported in Table 1 above are dialogues and the rest are monologues. However, there are several problems with this classification. For example, a debate, by its very name, assumes the presence of more than one speaker arguing over a specific topic. Moreover, while the main point of spontaneous commentary is to give a running commentary on an event, it often involves multiple speakers who are simultaneously engaged in a dialogue with each other. Similarly, court hearings may feature cross-examination of one person by another. Only prepared speech seems to fit the bill of a monologue. In our view, then, whether a text category is a dialogue or a monologue is a matter of degree rather than all-or-nothing. Furthermore, spontaneity is not only a feature of monologues but also dialogues; for example, interviews are considerably more prepared than face-to-face conversations. Finally, it is not only public discussions that are in the public domain; the same applies to all the monologues in LLC-2 that have been collected from public sources (see Section 5.2 below).

Thus, to determine the characteristic features of the text categories in Table 1 and the differences between them, we asked [number] people at Lund University to rate the degree of interactivity, spontaneity and privacy of each text category on a scale from 1 to 10 (e.g., how interactive, spontaneous and private is an *average* face-to-face conversation?). The respondents were given brief descriptions of the text categories and the three features. First, interactivity was associated with the number of people involved in a speech setting. Second, spontaneity was described as the extent to which the speech setting is written or practiced in advance of an event. Third, privacy was linked to the extent to which the speech setting is carried out in an intimate setting versus in front of an audience, either in person or via different media channels. Figure 1 presents the results of the survey in 3D space where spontaneity is given on the x-axis, privacy on the y-axis and interactivity on the z-axis.

[**Figure 1.** The arrangement of LLC-2 text categories in 3D space in relation to their level of interactivity (x-axis), spontaneity (y-axis) and privacy (z-axis)]

[Description of the figure.]

4. Comparisons with LLC-1

As mentioned above, LLC-2 is both a synchronic corpus and a comparable corpus to LLC-1. This section presents a general overview of LLC-1 (Section 4.1), and the similarities (Section 4.2) and differences (Section 4.3) between the two corpora.

4.1 LLC-1

LLC-1 was the first machine-readable spoken language corpus in the world. It grew out of the collaborative work between the Survey of English Usage at University College London, which was founded by Sir Randolph Quirk in 1959, and the Survey of Spoken English, founded by Jan Svartvik in 1975 as a sister project of the Survey in London. The main reason for compiling LLC-1 was to provide resources for a comprehensive description of the grammar of spoken English and to document the use of English in different types of speech settings for teaching and research purposes.[2]

LLC-1 comprises 100 texts of approximately 5,000 words each, totalling some 500,000 words for the whole corpus. The corpus data extend over four decades from the 1950s until the 1980s and include adult educated speakers of British English. The transcriptions in the corpus are orthographic and accompanied by detailed prosodic annotation, such as tone units, stress and pauses, but also simultaneous talk, contextual events and incomprehensible words. Each text is accompanied by information about the text type, year of recording, speaker age and so-called speaker category, which includes information about the speaker's gender and what can roughly be described as his/her occupation (but see below). The corpus is accessible for a small fee from the Survey of English Usage on CD-ROM as part of the Diachronic Corpus of Present-Day Spoken English and from the corpus management and analysis system

[2] See http://www.ucl.ac.uk/english-usage/archives/seu-biblio.htm for a complete list of publications based on work on the corpus. The link also contains publications related to the compilation of LLC-1 to which the reader is referred for more information about the corpus than can be provided here.

Corpuscle developed at the CLARINO Centre Bergen in collaboration with the University of Bergen, Norway (sign-up required). The online interface relies on the XML formatting of the corpus, but the XML files themselves have not been released to the public. The same applies to the sound files, which have not been anonymised and therefore cannot be made publicly available.

Figure 2 presents a partial design of LLC-1.

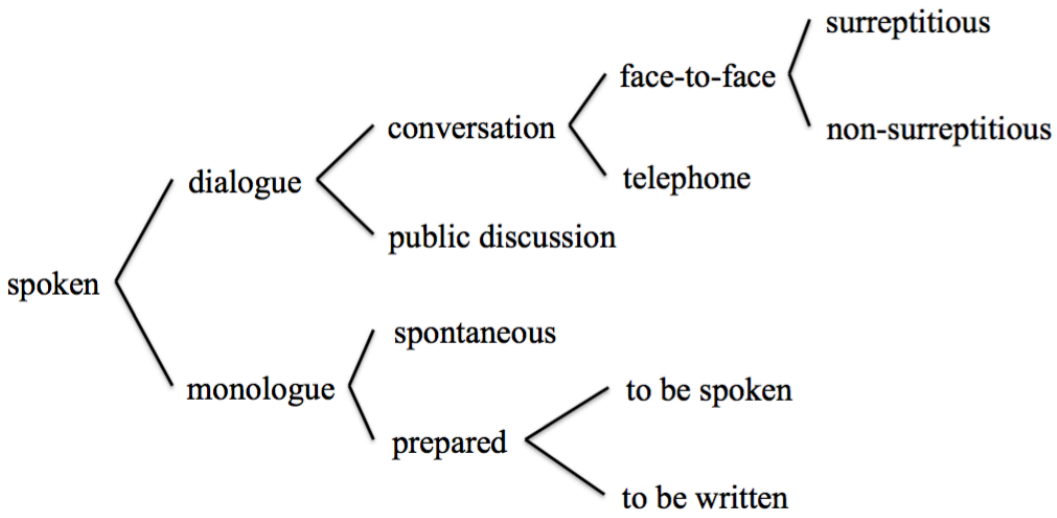

Figure 2. Partial design of LLC-1.[3]

As can be seen in Figure 2, LLC-1 includes both monologue and dialogue. Dialogues are further divided into conversations and public discussions. The former include both face-to-face conversation, where the speakers can see each other, and telephone conversations, where the speakers are not in the same place. All the telephone conversations, and most of the face-to-face conversations, were recorded surreptitiously, meaning that at the time of the recording one or more speakers in the conversation were not aware of it. Dialogues may also be public discussions, which are broadcast (mainly) on the radio and involve an audience. Monologues are either spontaneous or prepared. Spontaneous monologues are relatively unplanned, while prepared monologues are closer to written English. The latter category is further divided into speech that is to be spoken and speech that is dictated and therefore to be written down.

4.2 Similarities

LLC-1 and LLC-2 are comparable corpora in the sense that they differ from each other in terms of only one parameter, the parameter of time. The data in the former were recorded

[3] The figure has been adapted from Greenbaum & Svartvik (1990, p. 13).

between the 1950s and the 1980s, and the recordings in the latter were made between 2014 and 2019. While the time span in LLC-2 is relatively narrow, i.e., only five years, LLC-1 was recorded across four decades and therefore covers a much longer time span than LLC-2. Still, only a small number of the recordings in LLC-1 were made in the 1950s and the 1980s, which means that the bulk of the data come from the two decades in between. It is therefore safe to say that approximately 50 years separate the two corpora.

The rest of the parameters have been kept constant to the extent possible. Both corpora contain recordings of spoken British English involving adult educated speakers of the variety. Most of the speakers in face-to-face conversations and (mobile) phone/Skype conversations are associated with the University College London either through work or study. Importantly, the size and design of LLC-1 were taken to be the ultimate goal in the compilation of LLC-2 and, by and large, the goal has been achieved. First, the size of LLC-2 corresponds to that of LLC-1, with roughly 500,000 words in both corpora. Second, LLC-2 covers all the text categories in LLC-1 as discussed in Section 3 above. Nonetheless, there are minor differences within the categories themselves, largely due to what Leech (2007) calls 'genre evolution'. For example, in order to better represent the communication and broadcasting technologies used in the 21st century, landline telephone calls in LLC-1 have been replaced by mobile phone calls and video calls carried out over Skype. Moreover, broadcasting is no longer restricted to two media channels, TV and radio, but also includes webcasting and media distributed over the internet. As a result, broadcast discussions and interviews in LLC-2 include podcasts, and spontaneous commentary includes commentary and demonstrations published on YouTube.

Another important point of reference in the compilation of LLC-2 was the proportion of the text categories in LLC-1. Therefore, efforts were made to match the number of texts and words in each category as closely as possible. The proportions of the seven text categories in the two corpora are presented in Table 2. Due to intra-categorial differences between the corpora, only the subcategories of face-to-face conversation are given. The corpus user is advised to consult the LLC-1 corpus manual, together with the information provided in Table 1 above, for comparisons between the other subcategories. [Description of the table.]

Table 2. Comparison of the proportions of text categories in LLC-1 and LLC-2 in terms of the number of texts, words and speakers.

Text category	**Subcategory**	**LLC-1**			**LLC-2**		
		#T	#W	#S	#T	#W	#S
Face-to-face conversation	Equals	[…]	[…]	[…]	[…]	[…]	[…]
	Disparates	[…]	[…]	[…]	[…]	[…]	[…]
(Mobile) phone/Skype conversation	NA	[…]	[…]	[…]	[…]	[…]	[…]
Broadcast discussions and interviews	NA	[…]	[…]	[…]	[…]	[…]	[…]
Parliamentary language	NA	[…]	[…]	[…]	[…]	[…]	[…]
Spontaneous commentary	NA	[…]	[…]	[…]	[…]	[…]	[…]
Legal language	NA	[…]	[…]	[…]	[…]	[…]	[…]
Prepared speech	NA	[…]	[…]	[…]	[…]	[…]	[…]
		[…]	[…]	[…]	[…]	[…]	[…]

Note: '#T' refers to the number of texts, '#W' to the number of words and '#S' to the number of speakers.

4.3 Differences

Despite the similarities above, LLC-1 and LLC-2 also differ from each other in a number of ways. In what follows, three main differences are briefly outlined. The first one concerns the extent to which metadata, both about the texts and the speakers, were collected and made available. As mentioned above, in LLC-1 speaker metadata are limited to age, gender and occupation, but even these demographic categories are problematic. For example, age is often given as an estimate rather than an exact number, and occupation is sometimes viewed as the role taken on by the speaker in a given situation rather than his/her role in society. Among academics, businessmen and politicians, for example, one text in LLC-1 features a 'cookery expert' and a 'vegetarian'. Naturally, the term 'vegetarian' says nothing about what the speaker does for a living. There is also some degree of uncertainty about the time at which the recordings were made as many dates are given as estimates. By contrast, LLC-2 provides detailed information about the speakers, which has been obtained through questionnaires filled in by the speakers themselves or by careful online searches of public figures. In addition to age, gender and occupation, the questionnaire also elicited information about education, (foreign) language use, place(s) of residence and accent (see Section 6 below). Moreover, the speakers were asked about the exact date of the recording. Thus, the availability of comprehensive metadata in LLC-2 allows for more sophisticated sociolinguistic analyses of the data than in LLC-1.

The second major difference relates to whether or not the speakers in the recordings were aware of being recorded. As mentioned above, many of the speakers in LLC-1 were unaware of being recorded and were given no reason to deviate from their usual speech behaviour. The ethical guidelines and regulations of the 21st century, however, do not allow for recordings without prior consent, which means that the conversations in LLC-2 may not be as natural as the ones in LLC-1. The speakers' awareness of the unnatural circumstances of the speech setting is occasionally manifested in their comments on the recording equipment, the research project or their uptake of topics that are clearly primed by the recording situation such as the English language and Sweden. To prevent these comments from skewing the corpus results, obvious references to the recording have not been transcribed (see Section 7 below).

The third difference between the two corpora lies in the transcription and markup conventions used. Particularly, the transcriptions in LLC-2 do not follow the same principles as in LLC-1. Instead, they are a combination of the XML markup language and the transcription scheme of the International Corpus of English. The main reason for using a

different transcription and markup scheme in LLC-2 was to ensure compatibility with modern corpus tools (see Section 7 below for details). Another difference between the corpora is that the detailed prosodic annotation in LLC-1 has been discarded in LLC-2, largely due to time constraints. Instead, all the transcriptions in LLC-2 are accompanied by sound files, which allow corpus users to undertake additional analyses of the data themselves. No such sound files are available for LLC-1.

It is of utmost importance that the researchers who use LLC-1 and LLC-2 for diachronic analyses are aware of the differences between the two corpora. We have outlined the main differences here, but the corpus user is encouraged to consult the corpus manuals of both corpora before undertaking any diachronic analysis of the data. Yet, despite these differences, we believe that a satisfactory level of comparability between the corpora has been achieved.

5. Data collection

The data collection for LLC-2 was largely opportunistic. While the design of LLC-1 provided clear criteria as to which text categories should be targeted and to what extent, no such criteria existed for demographic categories such as age, gender and occupation. Therefore, in collecting data for LLC-2, we did not actively seek out recordings from certain groups of speakers but instead accepted recordings from all people interested in contributing to the project. At a later stage, when it became clear that some demographic categories were skewed towards certain groups of people, efforts were made to remedy the disproportions. One of the ways in which this was achieved was by encouraging the respondents[4] to involve certain groups of speakers in the recordings. For example, since it was easier to recruit younger rather than older people, the former were asked to involve speakers of their parents' or grandparents' age instead of recording with people of their own age. It was easier to target specific demographic groups during the collection of public data, largely thanks to the wealth of data found online (see Section 6 below for the distribution of the demographic categories in LLC-2).

The data for LLC-2 were collected in two parts over a period of five years, 2014–2019. The first part was concerned with recordings of private speech at three universities in the UK and Sweden (Section 5.1), and the second part involved data collection from various public

[4] In what follows, we will use the term 'respondent' to refer to people who were responsible for making the recordings, and the term 'speaker' to refer to all the other people in the recordings.

sources on the internet (Section 5.2). The following sections provide an overview of the recruitment and recording procedures, considerations of ethics and copyright laws and the collection of speaker metadata.

5.1 Private speech

Recordings of private speech in LLC-2 were advertised and carried out in three universities: University College London and Lancaster University in the UK and Lund University in Sweden. The University College London and Lancaster University were the sites of recording of face-to-face conversation and mobile phone/Skype conversation, and university lectures, which are a subcategory of prepared speech, were recorded at Lund University. We start by describing the process of collecting data of face-to-face and mobile phone/Skype conversation.

Recordings of face-to-face and mobile phone/Skype conversation were advertised via posters and leaflets on and near the university campus, on online mailing lists and social media pages and the departments' administrative offices. In London, posters were also put up at other universities, for example, in Queen Mary, Goldsmiths and City, all three part of the University of London. Many of the respondents were recruited through networking.

First, the respondents were given the option of recording either a face-to-face or a mobile phone/Skype conversation. If they chose to record a face-to-face conversation, they were asked if they preferred to record the conversation themselves or have it recorded by one of the corpus developers, namely Nele Põldvere (henceforth NP). A short meeting was set up with those who chose the first option to hand over the equipment and give detailed instructions. It was emphasised in the meeting that the speakers may choose to talk about any topic they like and that the conversations should not involve more than five speakers. Moreover, the recording was to be at least 35 minutes long so as to reach the 5,000-word limit as described above. The respondents were then given a sufficient number of copies of consent forms and questionnaires to be distributed among the speakers (see below). After the recording, another meeting was set up where the respondents returned the equipment and provided additional information about the recording (e.g., the date and place, the speakers in the recording, their relationship). Most conversations among equals in the corpus were recorded in this way. Conversations among disparates, however, were mainly recorded by NP. Importantly, NP was not present in the room when the recording was made.

University lectures at Lund University were recorded in a similar way to conversations among disparates. Departmental offices at the university were contacted to reach speakers of

British English in Lund who lecture. After establishing contact with the respondents, NP went to the designated location, recorded the lecture and took care of the documentation. A small number of the respondents submitted pre-recorded lectures used for flipped classroom teaching.

Both face-to-face conversations and lectures were recorded with the Zoom H4n Handy Recorder and were saved as uncompressed Waveform (WAV) audio files, with sample rate 44,100 times per second with 16 bits per sample. In the case of university lectures where the respondents often moved around the room, the recorder was connected to the external microphone Shure MX393/O. When the recording involved large groups of speakers or speakers whose voice qualities were similar, the conversations were video-recorded with the Denver AC-1300 action camera. The video recordings were only used for identification purposes during transcribing and will not be released to the public.

If the respondent chose to record a mobile phone/Skype conversation, a different procedure was followed. First, the respondents were given general information about the recording via email (e.g., the topic of conversation, preferred number of speakers, recording length). They were also given two options: to record a mobile phone call or a Skype video call. Internet voice calls were also considered, but unfortunately there are no programs available for recording VoIP software applications such as Viber and WhatsApp, and so no such recordings were made. In the case of mobile phone calls, the respondents were instructed to download a free app for recording calls (e.g., Call Recorder), and Skype video calls were recorded by providing the respondents with the appropriate software; Mac users were sent the link to download *Call Recorder for Mac*[5] and PC users received a key to download *SuperTintin Skype Video Recorder for PC*[6]. The mobile phone calls were saved in WAV or MP3 format and the video calls were saved as MP4 video files, which were later converted to WAV to remove the video and thereby protect the anonymity of the speakers. All the recordings were uploaded to a private folder on LU Box, which is a secure file sharing and data storage platform, and additional information about the recording was collected via email. Because the mobile phone/Skype conversations were recorded at a later stage of the project and under time constraints, the respondents were offered compensation in the form of one £10 Amazon gift voucher.

[5] https://www.ecamm.com/mac/callrecorder/

[6] https://www.supertintin.com

It should be noted that some of the face-to-face conversations in LLC-2 were recorded in a similar way to mobile phone/Skype conversations. In particular, this was the case in later stages of the project when data collection in the UK had come to an end, but when people still expressed interest in taking part in the project. In this case, the respondents were asked to record the conversation using a voice recording app pre-installed on their smartphone, save it in WAV or MP3 format and upload it to LU Box. Since the quality of a smartphone recording is generally lower than that of a digital voice recorder, the respondents were encouraged to record the conversation in a (relatively) quiet environment and with limited movement of the speakers. They were also asked to limit the number of speakers in the conversation to three.

All the speakers in the recordings were required to sign a consent form (see Appendix B). The consent form contained general information about the research project and the task, a statement about the speakers' rights, and the subsequent management and dissemination of the data. Moreover, by signing the consent form, the speakers agreed to make Lund University the copyright owner of the material. The respondents who submitted pre-recorded lectures remained the copyright owners of the recordings, but granted the corpus developers unrestricted use of the material. The students in the lectures were not required to sign a consent form because their contributions were deleted from the sound files, but they were informed of the recording beforehand and given the opportunity to approach NP in case of any objections. The same principle was used with people who were not part of the conversation proper (e.g., a waiter in a cafeteria). Importantly, the consent forms were signed by hand. The respondents who did not meet NP in person (e.g., who recorded a mobile phone/Skype conversation) were asked to sign a printout of the consent form, scan it and return it by email.

Besides the consent form, all the speakers in the recordings also filled in a questionnaire (see Appendix C). This was done either by hand or electronically on SurveyMonkey, which is a free online survey tool. The contents of the questionnaire are discussed in detail in Section 6 below.

5.2 Public speech

Recordings of public speech in LLC-2 were extracted from various internet sources. First, relevant material on the internet was identified and, if possible, downloaded in WAV or MP3 format. Relevant here means that (i) it was possible to classify the recording as one of the text

categories in the corpus design, (ii) the recording was made or published[7] no earlier than January 2014, and (iii) all the speakers in the recording were from the UK. Next, the perceived copyright owner(s) of the material were contacted by email and unrestricted use of the material, both in written and auditory format, was requested. The copyright owners were assured that the material would only be used for the purposes of non-commercial research and teaching. If the recordings were not available for download, copies of the files were also requested. Requesting copyright permissions proved to be one of the most challenging parts of the project, because in most cases the copyright owners simply did not reply to our requests. The ones that did reply did so in one of three ways: (i) the copyright permission was granted immediately by way of an informal, written statement (e.g., podcasters, YouTubers), (ii) the corpus developers were asked to sign a licence agreement, which granted the copyright permission on the conditions specified in the agreement (e.g., Houses of Parliament, Supreme Court), or (iii) the request was rejected due to the absence of a licence agreement to licence material for non-commercial use (e.g., TED Talks).

In most cases, the recordings of public speech were not collected directly from the speakers in the recording, which means that it was not possible to collect detailed demographic information about the speakers in the same way as for private speech. Efforts were made to gather basic information about the speakers from the internet, but in order not to risk reporting false information, we limited the information to only a few demographic categories. Section 6 provides more information on this topic.

6. The speakers

There are [number] speakers in LLC-2. The speakers are adults, i.e., they are above the age of 18, and all of them have completed at least upper secondary education for pupils aged between 17 and 18 years (the British equivalent of A Level). The only exceptions are two speakers who are 16 and 17 years old with the former not having completed A Level at the time of the recording.[8] Moreover, most of the speakers in the corpus are from England, and

[7] In some cases, there was no way to determine the exact date of the recording without asking the copyright owners. Since we wished to keep the process of requesting copyright permissions as brief and simple as possible, the date of recording for these recordings is the date when the recording was first released to the public.

[8] Although the respondents were asked to only record speakers who were above the age of 18, the metadata collected from these two speakers revealed that they were not. However,

only a small number of them come from other constituent countries of the United Kingdom (Wales, Scotland, Northern Ireland). Their first language (or one of them) is British English, and most of them consider themselves to have a British English accent of some sort (see below). A small number of the speakers are speakers of another variety of English (e.g., Irish, Australian) or another language altogether; however, in this case they had lived in the UK for a considerable period of time (>10 years). The contributions of these speakers have been marked with NN (non-native) in the speaker ID. In contrast to LLC-1, however, the contributions have been added to the total word count of the corpus to reflect the increasingly multicultural and multilingual landscape of the UK over the last few decades.

As mentioned above, the extent to which metadata about the speakers in LLC-2 are available depends on whether the recording is private or public. We will refer to the former group of speakers as Group A ([number] speakers) and the latter group as Group B ([number] speakers). While Group A provided detailed information about themselves, including age, gender, occupation, education, (foreign) language use, place(s) of residence and accent, only the first three demographic categories are available for all the speakers in Group B. This is because the rest of the information is more difficult to find online. In what follows, the distribution of the demographic categories in LLC-2 are presented one by one (Sections 6.1–7), and Section 6.8 compares the metadata available for both LLC-1 and LLC-2.

6.1 Age

Age is the first demographic category that is available for speakers in both Group A and B. In Group A, the speakers were asked to report their date of birth instead of current age. For some people, age may be a sensitive topic and in this way we hoped to elicit more responses. However, some speakers still did not answer the question, which forced us to determine their age based on the years they had reported to have lived in foreign countries and/or different locations in the UK (see below).

In order to allow for quantification and comparisons with other corpora such as the Spoken BNC2014 (Love et al., 2017), we divided the ages into four age groups: 16–34, 35–59, 60+, Unknown. The last group includes those speakers in Group B whose date of birth could not be found online. The four age groups are specific enough to illustrate generational differences between the speakers and broad enough to prevent data sparseness. In the

---

considering that the recordings also featured their parents as their legal guardians, we decided to keep the recordings in order not to lose valuable data.

metadata, age is available both as a number deducted from the date of birth (e.g., 30) and an age group. Figure 3 presents the distribution of the age groups in LLC-2. [Description of the bar chart.]

[**Figure 3.** The distribution of age groups in LLC-2]

It should be noted that there are a few speakers in the corpus that appear in more than one recording and at different points in time. One such speaker is the Queen of the United Kingdom, Elizabeth II, who features in four annual speeches given in the House of Lords at the State Opening of Parliament over the period 2014–2017. In order to avoid inconsistencies in data analysis, we decided to report the age of these speakers at the time of the first recording.

6.2 Gender

The speakers in Group A were given three options for gender classification: Male, Female and Other. In the questionnaires returned to the corpus developers, only the two first options had been marked. Moreover, the speakers in Group B have not given us any reason to believe that they identify with a gender other than male or female. Therefore, the bar chart in Figure 4 presents the distribution of Male and Female but not Other.

[**Figure 4.** The distribution of gender in LLC-2]

[Description of the bar chart.]

6.3 Occupation

In the questionnaire, the speakers in Group A were asked to write down their occupation in freeform text. The speakers in Group B were either public figures or they were recorded while carrying out professional duties (e.g., a preacher in a sermon), which made it possible to determine what they do for a living. In order to facilitate quantification, the occupations were mapped onto the social grade categories of the National Readership Survey's Social Grade demographic classification system.[9] Table 3 presents the categories of the Social Grade

[9] See http://www.nrs.co.uk/nrs-print/lifestyle-and-classification-data/social-grade/.

system. The metadata list both the social grade and the freeform text provided by the speakers.

Table 3. The codes and descriptions of the National Readership Survey's Social Grade.

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

Figure 5 presents the distribution of the speakers' occupation according to the categories of the Social Grade system. [Description of the bar chart.]

[**Figure 5.** The distribution of social grade in LLC-2]

Another possible estimate of the speakers' socio-economic status is the National Statistics Socio-economic Classification, which has been implemented in the Spoken BNC2014. This classification system is more nuanced than the Social Grade System, and it is also the government standard in the UK census. The reason for not using it in LLC-2 is because the fewer categories in the Social Grade system conform better to the smaller sample size of LLC-2 where there is a higher risk of data sparseness.

6.4 Education

Education is the first demographic category for which information is available only for Group A. It was possible to establish the educational level of some of the speakers in Group B, but the information is fragmentary and therefore not reported here. Where possible, it is reported in the metadata. In Group A, the speakers were given a pre-determined list of levels of education and asked to select the *highest* level that they had obtained by the time of the recording. The responses are summarised in Figure 6.

[**Figure 6.** The distribution of education in Group A in LLC-2]

[Description of the bar chart.]

6.5 (Foreign) language use

(Foreign) language use is concerned with what languages the speakers in Group A know (i.e., speak and/or understand) and when they started learning them. This information may be useful for explaining some of the non-English features used by the speakers or the possible attrition of their first language(s) due to influence from other languages. The average number of languages listed per speaker is [number], which means that it is common for the speakers in Group A to speak and/or understand more than one language. Table 4 lists all the languages reported by the speakers in the order of frequency, and the average age when they started learning them. [Description of the table.] In the metadata, (foreign) language use is given in the form of a list followed by the age of acquisition (e.g., English, 0; French, 12).

[**Table 4.** The languages spoken and/or understood by speakers in Group A in LLC-2 and the average age of acquisition]

The questionnaire also prompted the speakers to specify which of the languages listed is/are their native language(s) (marked with NL in the metadata). While the main reason for including this question was to confirm that the speakers were indeed eligible to participate in the project, it also gives interesting insights into multilingualism and linguistic diversity in the UK in the 21st century. While most speakers in the corpus ([number]) have one native language, [number] speakers have more than one native language. Table 5 lists the languages in the order of frequency.

[**Table 5.** The language(s) that speakers in Group A in LLC-2 consider to be their native language(s)]

[Description of the table.]

6.6 Place(s) of residence

In addition to listing the languages that they speak and/or understand, the speakers in Group A were also asked to list all the countries and towns/cities in the UK where they had lived for

longer than three months. The reason why it was decided to prompt the speakers to list all the places rather than the place of birth or current place of residence was because in this way we hoped to obtain a much more comprehensive picture of the role of geographical mobility in language use. On average, the speakers in Group A have lived in [number] different countries and [number] different towns/cities in the UK. This shows that, in addition to being multilingual, an average speaker in Group A had been exposed to different cultures and ways of living, both abroad and in the UK, which is likely to have had an impact on their use of English. The metadata contain all the countries and UK towns/cities listed by the speakers and the duration of the stay (e.g., USA, 5 months; London, 30 years). Table 6 presents the countries listed by the speakers together with the average time spent there, and Table 7 the UK towns/cities.

[**Table 6.** The countries where the speakers in Group A in LLC-2 have lived for an extended period of time (i.e., longer than three months) and the average time spent there]

[**Table 7.** The UK towns/cities where the speakers in Group A in LLC-2 have lived for an extended period of time (i.e., longer than three months) and the average time spent there]

[Description of the tables.]

6.7 Accent

Finally, the speakers in Group A were asked if they have a British English accent and, if so, which one. Since the main reason for asking this question was to further confirm their eligibility to participate in the project, no further attempts were made to verify the accuracy of the speakers' judgement, for example, against the recordings. Similar to the demographic categories described above, however, the question gives us important insights into dialectal variation. Since the answers were given in freeform text, they vary considerably in terms of both region and social class. For instance, the level of specificity of regional accent differs noticeably from one speaker to another. The speakers characterised their accent by country (e.g., Welsh, Scottish), supra-region (e.g., Southern, Midlands, Northern), administrative region (e.g., South East, West Midlands, North West), county (e.g., Kent, Somerset, Cumbria) or town/city (e.g., London, Hull, Sheffield, Bolton). In addition to region, accent was also associated with social class (e.g., RP, BBC, upper-middle class, cockney). In many cases, the two scales were combined to give a more accurate description: 'Standard Southern British

English', 'London educated working-class', 'BBC English with a hint of West Midlands', 'Southern middle class', 'RP mixed with Midlands/Manchester'. Furthermore, many of the speakers considered their accent to be a mix of different regional accents: 'mixed British', 'mixed (lived in the South, but born in the North + family lives in the North)', 'mix of Yorkshire-Lincolnshire-Southern', 'Hull accent which is now very weak', 'North-West Leicestershire modified towards Standard Southern British English'. Others considered themselves not to have a specific accent or found it difficult to identify it: 'possibly Southern?', 'no specific accent', 'normal accent', 'neutral'. One speaker even admitted to changing his accent relative to the interlocutor: 'Estuary English with friends otherwise RP'.

At first sight, these responses strongly defy a strict classification, and it is for this reason that the metadata contain the freeform responses provided by the speakers. However, in order to be able to make comparisons with other existing corpora, we grouped the responses based on the Nomenclature of Territorial Units for Statistics classification scheme (also used in the Spoken BNC2014). This means that the self-reported accents were coded according to a four-level scheme: global (e.g., UK), country (e.g., English), supra-region (e.g., South) and region (e.g., London). Undefined accents were coded as 'Unspecified'. Due to the fragmentary information provided by the speakers, e.g., when they specified the supra-region and not the region, the last level to be considered for quantification was the supra-region. Therefore, Figure 7 presents the distribution of the accents by supra-region only.

[**Figure 7.** The distribution of accent by supra-region in Group A in LLC-2]

[Description of the bar chart.]

6.8 Comparisons between LLC-1 and LLC-2

It was established above that the metadata available for LLC-1 and LLC-2 differ considerably. In LLC-1, the metadata include age, gender and occupation, but due to the uninformative nature of some of occupational titles in the corpus, it is not possible to map all of them onto the social grades used in LLC-2. This leaves us with age and gender. While gender is documented consistently in LLC-1, age poses further challenges, because it is often given as an estimate rather than an exact number. The broad classification of age groups in LLC-2, however, allows us to easily map the estimates in LLC-1 onto the four age groups as established above. Table 8 presents the comparison of age groups in LLC-1 and LLC-2, and Table 9 the comparison of gender.

[**Table 8.** The comparison of age groups in LLC-1 and LLC-2]

[**Table 9.** The comparison of gender in LLC-1 and LLC-2]

[Description of the tables.]

7. Transcription and markup

7.1 General information

The transcribing and marking up of the recordings in LLC-1 was carried out in two stages. This section is concerned with the decisions made in the first stage, i.e., the guidelines followed by the transcribers, and Section 8 below describes the additional markup and annotation of the transcriptions. A similar two-step procedure was adopted in the Spoken BNC2014. The transcription and markup scheme followed in the first stage was kept as simple as possible, largely in consideration of the workload of the transcribers. The transcriptions are orthographic, but the sound files allow the corpus user to extend the analysis of the corpus data beyond the orthographic transcriptions, for example, into the domain of prosody.

The transcriptions were produced following a detailed transcription and markup scheme. The scheme is largely based on a standardised markup language commonly used in corpus linguistics, namely XML (*eXtensible Markup Language*). XML works on the principle that whatever is enclosed within angle brackets is treated as corpus markup and whatever falls outside of the angle brackets is the actual corpus text. However, the markup language is not particularly human-friendly in the sense that inserting XML tags and all the attribute-value pairs associated with the elements in the tags is time-consuming for a human transcriber. Moreover, the XML standard does not specify how to best represent the textual component of the document, especially that of spoken language. We offer a few solutions. The first one relates to what has already been discussed above in relation to the two-step procedure of transcribing and marking up LLC-2, where in the first stage the transcribers manually inserted some of the XML tags, followed by a semi-automatic insertion of the rest of the tags in the second stage. Moreover, the tags inserted by the transcribers were very short and unobtrusive abbreviations of commonly used XML tags where only the first letter of the tag was included. For example, the tag for a pause was <p/> rather than <pause/>. The second solution was to use transcription software such as InqScribe (see below), which allows for a quick insertion of

pre-defined snippets, i.e., the XML tags. Finally, we combined insights from XML and corpus linguistics by following the transcription scheme of the International Corpus of English (ICE), which is specifically designed for spoken texts.

Therefore, LLC-1 and LLC-2 were transcribed and marked up using different conventions. There are two main reasons for why this was the case. The first one relates to the extent to which the conventions are known to the research community. The conventions in LLC-1 were only used in LLC-1 and have not been replicated in later corpora, whereas the XML standard is commonly used in corpus markup (see, for example, Hardie's, 2014 *Modest XML for Corpora*). Moreover, the ICE standard has formed the basis of more than twenty corpora of English worldwide, including the British component of ICE (ICE-GB), which together with LLC-1 forms the Diachronic Corpus of Present-Day Spoken English. The second argument against using the LLC-1 conventions in LLC-2 is age. LLC-1 was compiled before the advent of standard markup languages such as XML in the 1990s and is therefore incompatible with many of the corpus tools used today (e.g., Laurence Anthony's AntConc and Mike Scott's Wordsmith). For example, the LLC-1 transcription conventions include a variety of different symbols to represent, for example, simultaneous talk, *yes* or +yes+, contextual comments, (laughs), and incomprehensible words, ((yes)). Ideally, such representations are discarded in word counts because they do not constitute text proper. Modern corpus tools, however, are designed to discard text within angle brackets, which means that there is no straightforward way to discard all the symbols used in LLC-1. The ICE standard makes use of angle brackets, but not in the way that is compatible with the corpus tools. For example, in ICE contextual comments such as laughter are represented by <O>laugh</O>. AntConc would discard the letters 'O' but not the word 'laugh', which would then show up in word counts alongside the actual corpus text. As will be shown below, using a fully XML-compatible transcription and markup scheme as has been done in LLC-2 eliminates all the problems listed above.

7.2 The procedure

Prior to transcribing, the transcribers were given access to the transcription software InqScribe, a private folder in LU Box and three sets of documents: (i) general transcription and markup guidelines (see Appendix D), (ii) instructions on how to use InqScribe (see Appendix E) and (iii) two versions of the transcription and markup scheme, an extended version and a summary of the markup symbols (see below).

As mentioned above, the transcriptions in LLC-2 were carried out in InqScribe, which is low-cost digital media transcription software. A screenshot of the software is given in Figure 8. As can be seen in the figure, IncScribe has a simple interface for transcribing from both audio and video files. The sound control panel on the left-hand side allows the transcriber to start, pause, stop, rewind and forward the audio as well as manipulate the play rate. For further assistance, the Infinity IN-USB2 foot pedal was configured to the software to speed up the transcribing. The transcribing itself was done in the text box on the right-hand side. As mentioned above, InqScribe allows for the insertion of timestamps that link speaker turns in the transcriptions to the corresponding locations in the sound files. Another important feature of InqScribe is that it allows the transcriber to automatise the insertion of frequently used words and tags (called 'snippets' in the software), which means that, prior to transcribing, each markup symbol in LLC-2 was translated into a snippet and assigned a simple and intuitive shortcut on the keyboard (e.g., Shift-P for pauses). Finally, the software facilitates the exporting of the IncScribe file (.inqscr) in a variety of formats. The transcriptions in LLC-2 were exported as plain text files (.txt).

[**Figure 8.** Screenshot of a transcription in InqScribe]

Most of the transcriptions in LLC-2 were transcribed by one person, NP, who is a trained linguist. However, the majority of the face-to-face conversations, the core text category of the corpus, were transcribed by two people, one of whom is a native speaker of English. In this case, the transcribers did not work together but they worked separately on different transcriptions, which were then checked in full by the other transcriber. If a transcriber was in doubt about the accuracy of a word or a phrase, this part was marked up using a tag for unclear transcription as a signal to the other transcriber. More problematic cases were discussed further in person. If the transcribing was done by one person, the problematic cases were discussed with the other corpus developers. Either way, the transcriptions were proofread multiple times to ensure the highest possible level of accuracy. Despite these efforts, it is not possible to guarantee complete accuracy in the transcriptions, which may contain misses and inconsistencies that have escaped the transcribers' attention. It is therefore recommended that the corpus user works closely with the sound files linked to the transcriptions to detect any such errors. Moreover, we kindly ask corpus users who identify errors in the transcriptions to notify the corpus developers of them by sending an email to llc2@englund.lu.se.

7.3 Main features of the transcription and markup scheme

The transcription and markup scheme followed in LLC-2 is available for inspection in Appendix F (extended version) and Appendix G (summary). In what follows, the most important features of the scheme are presented and explained in order to further make transparent the decisions made in the process.

Speaker IDs and timestamps

Speaker IDs in LLC-2 are represented by 'S' followed by a unique numeric code. For example, the first speaker in the corpus is represented as <S001>. Speaker IDs were inserted together with timestamps, which specify the exact location of the speaker turn in the sound file. To facilitate easier retrieval of query searches in texts with only one speaker, a timestamp was inserted after every minute.

Segmentation by speaker turns

The minimal defined unit in LLC-2 is a speaker turn. No attempt was made to further divide the turns into orthographic sentences, as in ICE, or tone units, as in LLC-1. The reason for this is that both text units are notoriously difficult to identify and require a certain level of expertise by the transcribers. Speaker turns were deemed to introduce fewer errors and reduce the amount of work that goes into segmenting corpus texts. In the next stage, the speaker turns were marked and numbered (see Section 8.1 below).

Speaker turns may consist of one or more utterances. Example (1) illustrates three different types of speaker turns: (i) a long turn consisting of several utterances by speaker <S001>, (ii) a minimal turn consisting only of the backchannel *mhm* speaker <S002>, and (iii) a short turn by the same speaker.

(1) <S001> I think what's what would be useful now then now that you've kind of done the the quantitative <[>research</]> get a bit of qualitative <t>fe</t> kind of feedback from students who've been there we've got I think a couple of people with badges on who were at Melbourne
<S002> <[>mhm</]>
<S002> yeah I've I've spoken to people from Melbourne

In (1), it is the whole of <S002>'s first turn that is produced in overlap with <S001>'s previous turn, which means that it does not 'interrupt' it. This is not the case in example (2), where speaker <S004> produces the backchannel *oh good* slightly before speaker <S003> has finished her turn, resulting in an 'interruption' of the turn. Such an interruption was avoided in ICE in order to allow for syntactic parsing, but is not necessary in LLC-2 where the temporal unfolding of the conversation was considered to be more important.

(2) <S003> I think I'm gonna be off the last week of August that's
<S004> oh <[>good</]>
<S003> <[>what I was</]> thinking to use up that last week

The temporal unfolding of the conversation is also accurately represented in example (3). Here, speaker <S005> seeks confirmation of her statement that the preparations for welcoming new students are the same as the year before. The confirmation is provided by speaker <S006> with the word *same* to which <S005> swiftly responds with the backchannel *yeah* (*cause*). In order to accurately represent the causal relationship of the turns, speaker <S005>'s backchannel was inserted after <S006>'s turn and not at the end of her own prior turn.

(3) <S005> we're doing the same things as last year then <[>really</]>
<S006> <[>same</]> <[>yeah so</]> I mean the same uh the making the induction packs I'll draft the first-year lecture and seminar list
<S005> <[>yeah cause</]>

Pauses

Pauses are perceived breaks in conversation. The same tag (<p/>) was used to refer to both short and long pauses. The minimum length of a pause was considered to be one second, although this criterion was often overridden by the tempo of delivery. For example, what is considered a pause in face-to-face conversation is unlikely to be a pause in prepared speech. Ultimately, then, marking pauses was a subjective task and it is for this reason that corpus users interested in pauses (or gaps) are encouraged to also consult the accompanying sound files. Furthermore, it is important to emphasise the fact that text categories such as broadcast discussions and interviews that are often subject to editorial intervention commonly contain tags for pauses that do not represent natural breaks in conversation but rather the way in

which different bits of the programme have been put together in the editing process. Again, corpus users are encouraged to listen to the original recording for any cues that might help them differentiate between the two.

Note that the temporal unfolding of the conversation was also important in marking up pauses. This means that in examples such as (4), where there is a slight break between the two long turns by speakers <S007> and <S008>, the pause was inserted after the overlapping backchannel by <S008> rather than at the end of the prior turn, so as to communicate the fact that the pause was inserted immediately before <S008>'s long turn rather than after the overlap.

(4) <S007> what <[>we haven't</]> really done as a society is planned ahead and found the next set of antibiotics that replace the ones
<S008 > <[>yeah</]> <p/>
<S008> but what's interesting about this is as well is that obviously the bacteria are developing

Overlaps

In LLC-2, both the start and the end points of overlaps are marked. However, since overlaps are pervasive in spoken language, and particularly in spontaneous conversation, the insertion of XML tags for overlapping speech takes a considerable amount of effort. Therefore, the procedure of marking overlaps during the first stage of transcribing was considerably simplified. Following the ICE standard, we used the opening square bracket to signal the start of an overlap and the closing square bracket to signal the end of the overlap. Curly brackets were used to represent partial overlaps. Example (5) illustrates both types of overlaps.

(5) <S009> it makes me think about maybe crossing the strings rather than going up and down on the <[>fingerboard between the <t>s</t> <t>s</t> so <{>you of</}> a quick passage</]> you could play that way
<S010> <[>yeah yeah </{>yeah</}> that's right</]>
<S011> <{>yeah</}>

First, the example contains a long overlap between speakers <S009> and <S010> as represented by the square brackets. At the same time, parts of the turns are further overlapped by <S011>'s backchannel *yeah*. The curly brackets illustrate that the backchannel was uttered

at the same time as <S009>'s *you of* and <S010>'s *yeah*. In the next stage, both types of brackets were converted to conventional XML tags and numbered (see Section 8.1 below).

Filled pauses, backchannels and exclamations

Care was taken to represent filled pauses, backchannels and exclamations as accurately as possible. Following the ICE standard, filled pauses are marked as either *uhm* or *uh* depending on whether or not they end with a nasal sound. Backchannels are more varied. For example, backchannels expressing agreement range from the standard form *yes* to more conversational forms such as *yeah* and *yep*. They may also be expressed by *mhm* and *mm* depending on the presence or absence of the fricative. Furthermore, exclamations are represented by a number of different forms that speakers use to express either positive (e.g., *aw*, *yay*, *whee*) or negative emotions (e.g., *jeez*, *ugh*, *ow*). We believe that these sounds, alongside many others, serve important discourse functions in speech and need to be represented as precisely as possible. However, in order to be able to conflate very similar sounds onto fewer forms and thereby maximise corpus query recall, the transcribers were instructed to limit the sounds encountered in the recordings to a pre-defined list (see the transcription and markup scheme). If the transcribers did not succeed in matching the sound with any of the forms on the list, a tag was inserted instead (see below).

Non-verbal vocalisations

No pre-defined list was provided for marking up non-verbal vocalisations, such as laughing, coughing, yawning, blowing the nose and singing. That said, the transcribers were instructed to keep the number of different vocalisations to a minimum. The tag for non-verbal vocalisations is <v/> followed by a description of the vocalisation (e.g., <v desc="laugh"/>). The vocalisations that the transcribers failed to represent in a satisfactory way are described as 'other' (<v desc="other"/>). The same tag was used to represent unclassified filled pauses, backchannels and exclamations as described above. Note that no indication was made of the duration of the non-verbal vocalisation and the vocalisations that last for an extended period of time are only marked one time and at the start of the vocalisation. After transcribing, all the non-verbal vocalisations were extracted to establish the full range of vocalisations used in LLC-2. The following vocalisations were identified: [list of non-verbal vocalisations.]

Events

Events are only marked if they are contextually relevant and affect the comprehension of the interaction. Events include sounds, noises and actions of various kinds. They are marked up in a similar way to non-verbal vocalisations as described above but using the appropriate tag and description (e.g., <e desc="telephone rings"/>). Consider example (6), taken from a face-to-face conversation where one speaker calls in via Skype, and where a noise from the Skype recording interrupts speaker <S012>'s turn and prompts him to inquire about the source of the noise. The event affects the comprehension of <S012>'s turn and is therefore marked up in the transcription.

(6) <S012> if we do it in <e desc="noise from Skype"/> what's going on over there

The same markup is used to represent untranscribed text. There are two scenarios in which this was necessary. First, the text was not transcribed if the speakers commented on anything to do with the recording (e.g., equipment, confidential information, the time left to record), because the comments may create bias in the corpus results. Example (7) illustrates how this was done. In the example, speaker <S013>'s last turn is incomplete. The sound file, which was left unaltered, reveals that the speaker refuses to read out loud a text from his phone because of the recording. This part of <S013>'s turn has been concealed by the event tag containing the description of the event, 'comments on recording', which strongly implies that the reason for <S013>'s resistance to read the text out loud is due to his unwillingness to have it on tape.

(7) <S013> I'm going to read that in a second <v desc="laugh"/> <p/>
<S014> read it aloud
<S013> no <e desc="comments on recording"/>

The other reason for omitting text in the transcriptions was to protect the anonymity of the people who did not consent to being recorded, as shown in (8). In the example, speaker <S015> starts a turn that is interrupted by an unknown speaker who enters the room. Before leaving, the unknown speaker shifts the topic of the conversation to the Welsh national rugby team. Again, the markup conceals what the unknown speaker says, but it also explains why, after she has left the room, the speakers do not return to the earlier topic but continue talking about the Welsh team. For ethical reasons, the contributions made by unknown speakers have been deleted from the sound files.

(8) <S015> I reckon <e desc="unknown speaker enters the room, shifts the topic to the Welsh national team at the Rugby World Cup, leaves the room"/>
<S016> to be fair they have had a lot of injuries this year

Anonymisation

All the names and personal information that could potentially reveal the identity of the speakers in the recordings, and the identity of the people mentioned in the recordings, have been anonymised. The anonymisation involved marking up the personally identifiable information by enclosing it within the <a> and tags, and changing the information while retaining the word class and the number of syllables of the original. By not completely discarding the original text, we at least partly retain the socio-cultural information carried by a name, including gender and, in some cases, even ethnicity (Hasund, 1998). Consider example (9). In the example, speaker <S017>'s description of a friend reveals that the person being talked about is male and of Iranian descent. This information is reflected in the anonymised name, Ramin, which is a popular name given to boys in Iran. It is important to note that the above procedure does not apply to public figures.

(9) <S017> I don't know whether you've met <a>Ramin have you I suppose
<S018> I think <a>Dan did meet him in uh Weatherspoons several years ago
<S017> yeah he's Iranian American from California

Moreover, the anonymised name in (9) reflects the number of syllables of the original. This is in no way to compromise the anonymity of the speaker but to ensure that the written form matches the altered tonal pattern of the name in the sound file. The way anonymisation was done in the sound files is discussed at length in Section 8.3 below.

8. Additional markup and annotation

8.1 Additional XML markup

[To be written.]

8.2 POS-tagging and lemmatisation

[To be written.]

8.3 Anonymisation of the sound files

[To be written.]

9. Quick guide to using LLC-2

[To be written.]

References and links to relevant resources

AntConc: http://www.laurenceanthony.net/software/antconc/

CLARIN – European Research Infrastructure for Language Resources and Technology: https://www.clarin.eu

Corpuscle: http://clarino.uib.no/korpuskel/page?page-id=korpuskel-main-page

Greenbaum, S., & Svartvik, J. (1990). *The London-Lund Corpus of Spoken English*. Retrieved from http://clu.uni.no/icame/manuals/LONDLUND/INDEX.HTM

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.

InqScribe: https://www.inqscribe.com

International Corpus of English (ICE): http://ice-corpora.net/ice/

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer. (Eds.), *Corpus Linguistics and the Web* (pp. 133–149). Rodopi, Amsterdam.

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

Lund University Humanities Lab's corpus server: https://www.humlab.lu.se/en/facilities/corpus-server/

Svartvik, J., & Quirk, R. (Eds.). (1980). *A Corpus of English Conversation*. Lund, Sweden: Lund University Press.

The Diachronic Corpus of Present-Day Spoken English (DCPSE): http://www.ucl.ac.uk/english-usage/projects/dcpse/index.htm

The Spoken BNC2014: http://corpora.lancs.ac.uk/bnc2014/

Wordsmith: https://www.lexically.net/wordsmith/

List of appendices

Appendices

Appendix A. End User Licence agreement

The London-Lund Corpus 2 (LLC-2) of spoken British English

End User Licence agreement

Please read the licence carefully and make sure that you have understood all the terms and conditions. By downloading LLC-2, you agree to the terms and conditions specified here. Failure to comply with them results in the termination of the licence.

Terms and conditions

- LLC-2 must be used for the purposes of **non-profit research and teaching only**.
- LLC-2 must not be used for commercial purposes.
- Copyright in all LLC-2 texts and sound files is retained by the original copyright holders.
- The Licensee must not redistribute the password to the corpus download to any third party.
- The Licensee must not reproduce or redistribute all or any parts of LLC-2. A copy of the corpus may be made for backup purposes.
- The Licensee must use the corpus data in a way that is permitted under the fair dealings provision of copyright law. This means that the Licensee is only allowed to quote short passages from the corpus (up to 200 words from any given text).
- The Licensee must acknowledge LLC-2 in any publication arising from the use of the corpus data (see the reference below).
- The LLC-2 developers should be informed of any publication where the corpus data have been used (see the email below).

Sample reference:

- Põldvere, N., Johansson, V., & Paradis, C. (2019). The London-Lund Corpus 2 of spoken British English. Lund, Sweden: Lund University. Available from https://corpora.humlab.lu.se/

Please fill in the form below (electronically or handwritten in block letters). Your data will be used in accordance with the Data Protection Act.

Name: _____

Institution: _____

Postal address: _____

_____

Email address: _____

Date: _____

Signature: _____

LLC-2 is password-protected. To obtain the password, please send this agreement to **LLC2@englund.lu.se**. If filled in by hand, the agreement should be scanned and attached to the email in PDF format. The password will be sent at the developers' earliest convenience.

Appendix B. Consent form

CONSENT FORM

- **Title of project: The London-Lund Corpus 2 of spoken British English**
- **Name of researcher: Nele Põldvere**
- **Contact information: nele.poldvere@englund.lu.se**
- **Names of supervisors: Carita Paradis (carita.paradis@englund.lu.se) and Victoria Johansson (victoria.johansson@ling.lu.se)**

This project is concerned with the compilation of the London-Lund Corpus 2 of spoken British English. A spoken corpus is a collection of texts of spoken language in electronic format. The main objective of the project is to record and study spoken British English.

The project involves an audio and video recording of your conversation. Since we are interested in naturally occurring language, the topic of the conversation is of no importance to us and you can choose to talk about anything you like. Apart from the recording, the conversation should be as natural as possible.

The copyright owner of the recording will be Lund University. The audio recording will be transcribed and the files will be stored on the Lund University server. The analysis of the corpus data will be presented in a PhD thesis and other publications such as journal articles, and in academic conferences and seminars. Later, the corpus (both the audio recordings and the transcriptions) will be made publicly available from the Lund University Humanities Lab's corpus server for use in non-commercial research and teaching. The corpus will be password-protected and any user who wishes to gain access to the corpus needs to sign an End User Licence agreement. The video recording will only be

CONTINUE ON NEXT PAGE

used for identification purposes during transcribing by the corpus developers and will never be made publicly available.

The corpus will be fully anonymised, meaning that your name will never be connected to the recording. Instead, you will be given a unique ID number. All the names mentioned in the conversation will be changed in the transcription and distorted in the recording, making them impossible to identify.

Please note that participation in this project is entirely voluntary. You are free to refuse to participate and/or withdraw from the project at any point in time, but not later than 31 December 2018, and without giving any reason. Moreover, you can pause the recording if you need to discuss a sensitive topic or you can ask me to delete any part(s) of the recording afterwards (see contact details above).

By signing below, you are indicating your consent to be recorded for the project.

Today's date

…….

Name (please print)

…….

Signature

…….

Appendix C. Questionnaire

The London-Lund Corpus 2 of spoken British English

Recording ID:

Speaker ID:

QUESTIONNAIRE

Gender:

□ Male □ Female □ Other

Date of birth:

……

Occupation:

……

CONTINUE ON NEXT PAGE

Level of education:
(choose the highest obtained)

□ Doctorate (e.g. PhD, MD)
□ Postgraduate degree (e.g. MA, MSc)
□ Undergraduate degree (e.g. BA, BSc)
□ Vocational qualification (e.g. HNC, HND)
□ A-Level or equivalent
□ GCSE Level or equivalent
□ Other qualification (please specify):……...

Please list all the languages you know (speak and/or understand) in order of acquisition:
(native language(s) first)

	Language (write below)	**Native language**	**Age when you started learning**
1.	_____	□ Yes □ No	……. years old
2.	_____	□ Yes □ No	……. years old
3.	_____	□ Yes □ No	……. years old
4.	_____	□ Yes □ No	……. years old
5.	_____	□ Yes □ No	……. years old
6.	_____	□ Yes □ No	……. years old

CONTINUE ON NEXT PAGE

Please list all the countries where you have lived for an extended period of time (longer than 3 months):
(for UK also add the name of the town/city)

Country (write below)	**Total length of stay**	**City** (only for UK)
_____	……	……....
_____	……	……....
_____	……	……....
_____	……	……....
_____	……	……....
_____	……	……....

Do you consider yourself to have a British English accent? If yes, please specify which one:

……

COMPLETED BY THE RESEARCHER

Date and place:

……

Type of interaction:

……

File name:

…….....

Appendix D. General transcription and markup guidelines

- All the files are in your folder in LU Box. The folder has three subfolders: Documents, Recordings and Transcriptions. Documents contains the necessary documentation (see below), Recordings the sound and the video files to be transcribed and Transcriptions is where you will upload your transcriptions.
- The sound files to be transcribed always end with *cut.wav* or *cut.mp3*. Complete sound files have also been provided. Some recordings are accompanied by a video recording, which is useful in identifying the speakers.
- Please consult the InqScribe instructions (Documents) on how to use the transcription software.
- Before you start transcribing, check the information about the recording that you're about to transcribe and the speakers in it. Information about the recording can be found in recordings.xlsx and information about the speakers in speakers.xlsx. Both files are in Documents. Specifically, you should know the
    - recording ID,
    - speaker IDs,
    - speaker characteristics (gender, age, accent, etc.),
    - speech setting (face-to-face conversation, Skype conversation, etc.),
    - first words of the speakers.
- Start transcribing from the very beginning of the recording and the InqScribe file, and stop when you've reached 3,200 words for half texts and 6,200 words for full texts (see the last column of recordings.xlsx).
- When you transcribe, it is advised to first produce a rough transcript and then attend to details.
- The markup conventions used in the transcriptions are represented by tags surrounded by <angle brackets>. There are two types of tags. The first type takes scope over one or more words. In this case, the opening tag is <element> and the closing tag is </element>. Note that there is no whitespace between the brackets and the words within them. The second type of tag indicates a point rather than a region. In this case, it is represented by <element/> (note the placement of /). Some such tags contain attribute-value pairs (<element attribute="value"/>) where the attribute corresponds to the type of information and the value to the actual information.

- Always transcribe what you hear and not what you expect to hear.
- If you're not sure about a word or how to write it, try to google it.
- When you've finished transcribing, save the transcription in two formats (see the InqScribe instructions for details):
    - textID_yourname.inqscr
    - textID_yourname.txt
- Upload the two files to the folder Transcriptions after every transcribing session. You can use the form textID_yourname_01.txt for an unfinished version and textID_yourname.txt for the final version. Make sure to check the transcription several times before uploading the final version.
- To keep track of your transcriptions, regularly update the Progress Report (Documents). An example of how to do it has been provided for you in the file.
- All comments, questions, problems and suggestions that arise during transcribing should be written down in Comments.boxnote (Documents). For example, you may want to draw attention to a transcription that is finished but needs to be checked by another transcriber.

Appendix E. Instructions on how to use InqScribe

- Open InqScribe.
- To load a new sound file, open Media -> Select Media Source. Select the sound file that you want to transcribe and set Timecode to 00:00:00.00 by clicking Start at Custom Time. Click OK.
- Keep the default transcription settings (Helvetica 16).
- Set up the foot pedal by clicking Edit -> Set Up Foot Pedal. Assign a command to each pedal according to your liking. For example, you may want to assign Play until Released to the centre pedal (with an option to skip a number of seconds once released), Review to the left pedal and Jump to Beginning to the right pedal (useful when you've just started transcribing). You can change the commands by clicking Edit -> Edit Shortcuts. You can make the shortcuts list available while transcribing by clicking Window -> Show Shortcuts.
- Go to Edit and choose Edit Snippets to set up timestamps and speaker IDs. For example, if the speaker ID is <S001>, click Add, then go to Define Trigger and press any key that you want to use for that speaker (e.g., Command-1 depending on the speaker number). Click OK. Then highlight the text box below and click Insert Snippet Variable. Choose {$time} from the list. The snippet variable shows up in the text box. Enclose the variable within angle brackets, and insert a space, the speaker ID within angle brackets and another space after it: <{$time}> <S001>. Repeat the procedure for every speaker in the sound file. Finally, click Done. You can make the snippets list available while transcribing by clicking Window -> Show Snippets. Remember that you can't use the same key for more than one snippet.
- Start transcribing by pressing the foot pedal. Press the assigned snippet key at the start of each speaker turn to insert timestamps and speaker IDs. Write what the speaker says and press Enter at the start of the next turn. Do the same for all other speaker turns. Remember to pay close attention to the transcription and markup scheme and the general guidelines.
- Feel free to set up other snippets and shortcuts. You may want to use snippets for some of the transcription and markup tags (pauses, overlaps, non-verbal vocalisations) or commonly used sounds (*yeah*, *okay*, *uhm*).

- Remember to regularly check Transcript -> Word Count. When you've reached 3,200 words for half texts and 6,200 words for full texts, stop transcribing.
- Always save the transcription in two formats: .inqscr and .txt. To save the InqScribe file, click File -> Save and give it the following name: textID_yourname.inqscr. To save a plain text file, click Plain Text -> Export and give the file the same name.
- If you want to continue working on an already existing transcription, simply open the InqScribe file.

Appendix F. Transcription and markup scheme

Category	Definition and description	Symbols and examples
Text IDs	• Text IDs uniquely identify a text • They are marked at the very beginning and very end of each transcription file • The value always has two components: the letter 'T' and a unique numeric code (e.g., T001). • The text IDs can be found in recordings.xls	<text id="T001"> (beginning) </text> (end)
Speaker IDs	• Speaker IDs uniquely identify a speaker • They are marked at the beginning of each speaker turn, including backchannels • They have two components: the letter 'S' and a unique numeric code (e.g., S001); non-native speakers have an additional component: NN (e.g., S002NN) • They can be found in recordings.xls	<S001> this is a transcription
Timestamps	• Timestamps mark the beginning of each speaker turn • They are inserted before the speaker ID (see InqScribe instructions on how to do it) • In recordings with only one speaker, a timestamp is inserted after every minute	<[00:00:00.00]> <S001> this is a transcription
Speaker turns	• The transcriptions are segmented by speaker turns • Speaker turns are continuous stretches of speech produced by one speaker • Each speaker turn starts on a new line	<S001> this is the first speaker turn <S002> this is the second speaker turn
Utterances	• Speaker turns consist of one or more utterances	<S001> this is the first utterance this is the

	• Utterances are basic units of communication • They always begin with a lower case letter (but see Capitalisation below for exceptions)	second utterance
Punctuation	• No punctuation (e.g., full stops, question marks, exclamation marks) is used to indicate statements, questions or exclamations	I live in London (statement) what are you doing (question) I just won the lottery (exclamation)
Pauses	• Pauses are perceived breaks in conversation that are *usually* longer than one second • They are always marked within or at the end of a speaker turn but never at the beginning • If a pause occurs within a word, it is inserted *after* the word • The speaker's tempo of delivery must be taken into account	<p/>
Overlaps	• Overlaps are points at which more than one speaker is speaking • When the overlap occurs between two speakers, square brackets are used around the text that overlaps • When the overlap occurs between three or more speakers and the overlapping parts are identical in length, square brackets are used around the text that overlaps • When the overlap occurs between three or more speakers and only parts of the text overlap, curly brackets are used around the text that overlaps	<S001> it's a nice day <[>today isn't it</]> <S002> <[>yeah it is</]> <S001> it's a nice day <[>today isn't it</]> <S002> <[>yeah it is</]> <S003> <[>I agree</]> <S001> it's a nice day <[>today <{>isn't it</}></]> <S002> <[>yeah <{>it is</}></]> <S003> <{>mhm</}>
Filled pauses, backchannels and exclamations	• Filled pauses, backchannels and exclamations are all spelt out • The sounds are limited to the forms on the right • If the sound does not fit any of the forms given here, a tag is inserted	uh, uhm yes, yeah, yep, mhm, mm no, nope, nah

	instead (see Non-verbal vocalisations below)	okay oh, ooh, ah, huh, aha hey aw, oops, wow, phew, jeez, duh, yay, ugh, whoa, whee, ow
Non-verbal vocalisations	• Non-verbal vocalisations include laughing, coughing, yawning, sneezing, blowing the nose, whistling, singing (without lyrics), etc. • Extended non-verbal vocalisations are marked only once at the start of the vocalisation • Non-verbal vocalisations are written in the base form and as concisely as possible • They are not limited to the forms on the right, but the number of different non-verbal vocalisations should be kept to a minimum • If the non-verbal vocalisation, or the sounds described above, cannot be easily described, use the 'other' value	<v desc="laugh"/> <v desc="cough"/> <v desc="sing"/> <v desc="other"/>
Events	• Events are noises, sounds and actions of various kinds that are contextually relevant and affect the comprehension of the interaction • They also include comments on the recording and contributions by unknown speakers • They are written as concisely as possible and, if possible, in third-person • They are not limited to the forms on the right, but the number of different events should be kept to a minimum • Events that occur between speaker turns are indicated at the end of the preceding turn and never on a separate line	<e desc="telephone rings"/> <e desc="browses computer"/> <e desc="comments on recording"/> <e desc="unknown speaker enters the room, talks about a football game, leaves the room"/>
Anonymisation	• Any personal information about the speakers that could potentially reveal their identity are marked and changed, including information	Martin Smith (original) <a>Norman Croft (marked and

	about those who do not take part in the conversation • The word class and number of syllables of the original are preserved • Replacements must be consistent within a text • Names of public figures and places are not marked or changed	changed) Two Bread Street (original) <a>Two Grub Street (marked and changed) BUT NOT I went to school in Manchester, Barack Obama
Truncated words	• Truncated words are words that are unfinished, for example, false starts and repairs • They are marked and transcribed as much as can be made out	it's a <t>corp</t> a spoken language corpus
Repetitions	• All repeated words are transcribed as many times as they occur	it's it's nonsense I I I didn't know yeah yeah yeah yeah
Abbreviations, acronyms and spelt out words	• Some standard abbreviations are used but without punctuation • Acronyms are written together in caps • Plural acronyms are immediately followed by a small 's' • Spelt out words have spaces in between letters	Mr, Mrs, Ms, Dr, BUT I went to the doctor (abbreviations) UK, NATO, PhDs (acronyms) it's spelled E L I O T (spelt out)
Non-standard speech	• Non-standard speech includes contractions, shortenings and non-standard grammatical forms • They are transcribed as they are spoken • However, the transcriber must be absolutely certain that a non-standard form is used • Contractions are limited to the types of contractions on the right; others should be written out • Shortenings include semi-standard merged words but not shortenings occurring solely due to fast speech • No attempt should be made to correct uncorrected slips of the tongue	I'm, you'd, she'll, he's, we'd, they've, isn't BUT NOT might'd've, people'll've (contractions) cos, gonna, tryna, kinda, innit BUT NOT runnin, nothin, em (shortenings) could of been, I seen him, he don't, youse (non-standard grammatical forms)

Numbers	• Numbers are transcribed as they are spoken • This also applies to proper nouns (see Capitalisation below) but not to well-established acronyms • Hyphenation is only used between 21 and 99 • Number 0 pronounced as 'oh' is transcribed as 'o'	six, nineteen eighty-four, ten point five, nineteen o nine, seven thousand two hundred seventy-five, twenty-first, one o'clock Ten Downing Street BUT NOT 3D
Non-English words	• Words from languages other than English are marked and transcribed as they occur in the source language	<f>c'est la vie</f>
Capitalisation	• Only proper nouns and the personal pronoun *I* are capitalised	Sweden, University College London, Department of History (proper nouns) you and I (personal pronoun *I*)
Spelling	• Generally, British English standard spelling is used • Consult a dictionary if in doubt (e.g., Oxford, Cambridge, Longman, Collins, Macmillan) • If the spelling is used in both British and American English, google it and use the most popular form (also see the list on the right) • Use the spelling conventions on the right for hyphenated/non-hyphenated and other types of words	analyse, colour, centre, travelling, offence, focused, judgment (regional or popular spellings) so-called, non-binary, computer-related, email, reapply, substandard, cooperation, brownish (hyphenated/non-hyphenated) et cetera, per cent, encyclopedia (other)
Unintelligibility and uncertainty	• Unintelligibility and uncertainty are marked up using the same tag, which is applied differently • Unintelligibility is marked on words that are *completely* incomprehensible, often due to some external noise or extensive overlapping • Uncertainty is marked when the transcriber is in doubt about the accuracy of the transcription, including speaker turns	this is a <u/> (unintelligibility) this is an example of <u>radiation</u> filtering <u>S001</u> (uncertainty)

Appendix G. Summary of transcription and markup symbols

Text ID	<text id="T001">…</text>
Speaker ID	<S001>
Timestamp	<[00:00:00.00]>
Pause	<p/>
Complete overlap between two or more speakers	<[>…</]>
Partial overlap between three or more speakers	<{>…</}>
Non-verbal vocalisation	<v desc="laugh"/>
Event	<e desc="telephone rings"/>
Anonymisation	<a>Norman Croft
Truncated word	<t>corp</t>
Non-English word(s)	<f>c'est la vie</f>
Unintelligible word(s)	<u/>
Uncertain transcription	<u>radiation</u>