English Language and Linguistics, 25.3: 459–483.
doi:10.1017/S1360674321000186

On *The London–Lund Corpus 2*: design, challenges and innovations[1]

NELE PÕLDVERE
Lund University and University of Oslo
VICTORIA JOHANSSON and CARITA PARADIS
Lund University
(Received 18 November 2020; revised 20 May 2021)

This article describes and critically examines the challenging task of compiling *The London–Lund Corpus 2* (LLC–2) from start to end, accounting for the methodological decisions made in each stage and highlighting the innovations. LLC–2 is a half-a-million-word corpus of contemporary spoken British English with recordings from 2014 to 2019. Its size and design are the same as those of the world's first machine-readable spoken corpus, *The London–Lund Corpus of Spoken English* with data from the 1950s to 1980s. In this way, LLC–2 allows not only for synchronic investigations of contemporary speech but also for principled diachronic research of spoken language across time. Each stage of the compilation of LLC–2 posed its own challenges, ranging from the design of the corpus, the recruitment of the speakers, transcription, markup and annotation procedures, to the release of the corpus to the international research community. The decisions and solutions represent state-of-the-art practices of spoken corpus compilation with important innovations that enhance the value of LLC–2 for spoken corpus research, such as the availability of both the transcriptions and the corresponding time-aligned audio files in a standard compliant format.

Keywords: corpus compilation, spoken language, *The London–Lund Corpus of Spoken English*, XML transcriptions, open access

1 Introduction

The aims of this article are (i) to describe and critically examine the challenging task of compiling a spoken corpus from start to end, namely, *The London–Lund Corpus 2* (LLC–2) of spoken British English; (ii) to explain the methodological decisions that were necessary to make the corpus suitable both for synchronic investigations of

[1] The compilation of LLC–2 was largely funded by 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-Sörensen Foundation. It has also benefitted from an infrastructure grant from the Swedish Research Council (Swe-Clarin, 2019–2024; contract no. 2017-00626) and the support of the Lund University Humanities Lab. We are grateful to two reviewers and Laurel Brinton for their insightful comments on an earlier version of this article.

contemporary spoken language as well as for principled diachronic comparisons of speech with its predecessor, *The London–Lund Corpus of Spoken English* (LLC–1) with data from the 1950s to 1980s (Greenbaum & Svartvik 1990); and (iii) to highlight the innovations in the new corpus. Each part and every stage of the compilation of LLC–2 posed its own challenges for how to make it useful for a wide range of linguistic studies.

The importance of better knowledge about authentic spoken communication in the real world is self-evident; every day people participate in a range of contexts and situations as speakers, as addressees and as passive listeners. For example, we chit-chat with friends and family, make formal presentations at work, listen to the news and eavesdrop when other people talk. In other words, we both initiate and receive spoken communication in different forms all day long. In dialogic exchanges, opinions, ideas and viewpoints are exchanged very rapidly and smoothly; people act and respond on the spur of the moment. Meanings are constantly being negotiated to reach mutual understanding between interlocutors. Spoken corpora provide important insights into the kinds of expressions that speakers use in each exchange as well as into the kinds of processes that underlie human social behaviour, perception and cognition (Halliday 1989; Clark 1996; Garrod & Pickering 2004; Linell 2009; Du Bois 2014; Põldvere & Paradis 2019, 2020; Põldvere, Johansson & Paradis 2021).

Since spoken language and in particular collaborative impromptu speech in dialogic contexts is dynamic and in a constant flux, it also offers a fertile ground for the study of language variation and change. To better understand the motivations and mechanisms for this, there is a need for new spoken corpora at certain intervals in time for scientists of language and human behaviour to be able to monitor contemporary language use, and also for them to be able to make comparisons with language use and social interaction back in time. LLC–2 meets a long-felt need.

LLC–2 is a new half-a-million-word corpus of spoken British English with data from 2014 to 2019. It features 360 educated adult speakers from the UK, primarily from England, who appear in a variety of discourse contexts: face-to-face conversation, phone/CMC conversation,[2] broadcast media, parliamentary proceedings, spontaneous commentary, legal proceedings and, finally, prepared speech. It mirrors the size and design of LLC–1, but it also features the rapid technological advances of the twenty-first century, where, for instance, internet chats are commonplace. LLC–2 represents state-of-the-art practices of corpus archiving, distribution and preservation in that the transcriptions are released under open access for use in two different ways (as downloadable XML files and via an online interface). Moreover, the transcriptions are accompanied by corresponding time-aligned audio files to further enhance the value of LLC–2 for spoken corpus research (see Põldvere, Frid, Johansson & Paradis 2021 for details).

The procedure of this article is as follows. Section 2 provides an overview of LLC–1 and the main features of that corpus that prompted the compilation of LLC–2. Sections

[2] CMC = Computer-Mediated Communication.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

3–5 discuss the methodological decisions made in each stage: the design of the corpus (section 3), data collection, ethical and legal considerations, and the recruitment of speakers (section 4), and, finally, the transcription of the recordings, and markup and annotation procedures (section 5). Section 6 compares the main features of LLC–1 and LLC–2. The article ends in section 7 with information about how users can access LLC–2, and some concluding remarks.

2 *The London–Lund Corpus 1*

LLC–1 is the result of a collaborative effort between two projects at two institutions: the Survey of English Usage launched in 1959 at University College London by a team led by Sir Randolph Quirk and the Survey of Spoken English launched in 1975 at Lund University by a team led by Jan Svartvik. While a major undertaking of the Survey of English Usage was to collect and transcribe large amounts of recordings of authentic spoken British English to serve as a basis for studies of the nature of spoken language, the Survey of Spoken English took on the task of making available the spoken material in machine-readable form. The first copies of the computerised LLC–1 were distributed among interested scholars in 1980, making LLC–1 the first publicly available spoken corpus in the world.

LLC–1 is a half-a-million-word corpus of spoken British English recorded with educated adult speakers over a period of four decades, from the 1950s until the 1980s. Each text, 100 in total, is approximately 5,000 words in size. The computerised version of the corpus contains detailed prosodic transcriptions, annotated for features such as tone units, the location and direction of nuclear tones, pauses of different length, etc. The transcriptions also contain information about speaker identity, simultaneous talk, contextual comments and incomprehensible words. In the majority of cases, the year of recording and metadata about the speakers are given (see section 6.1 below for details). Access to the electronic corpus can be requested from two locations.[3] First, LLC–1 is accessible on CD-ROM upon payment of a licence fee from the Survey of English Usage as part of *The Diachronic Corpus of Present-Day Spoken English* (DCPSE; Aarts, Close & Wallis 2013). This release of LLC–1 contains approximately 400,000 words of spontaneous speech from the original corpus, fully tagged and parsed for parts-of-speech and dependency relations by the developers of DCPSE. The corpus can be searched via the ICECUP software (Wallis 2006). Second, the complete corpus can be accessed via the corpus management and analysis system Corpuscle, developed and maintained at CLARINO Centre Bergen in collaboration with the University of Bergen (Meurer 2012).[4] Similar to ICECUP, Corpuscle allows for the implementation

[3] But see also Svartvik & Quirk (1980) for a book version containing thirty-four texts (conversational transcriptions).
[4] CLARINO is the Norwegian member of the European Research Infrastructure for Language Resources and Technology (CLARIN). For more information on CLARIN, see www.clarin.eu

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

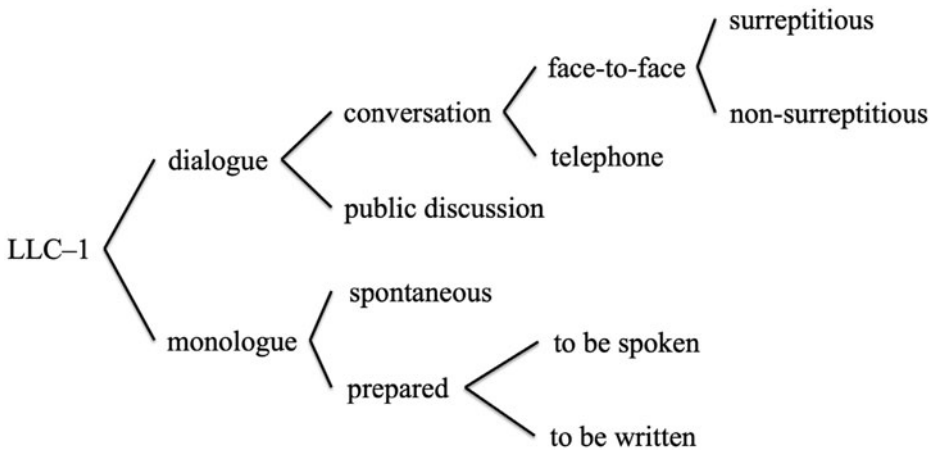

Figure 1. The basic design of LLC–1 (adapted from Greenbaum & Svartvik 1990: 13)

of various corpus linguistic techniques on LLC–1 such as query searches, and concordance and collocation analyses, but in contrast to DCPSE, access to the corpus is free via institutional login (see section 7 below for details on Corpuscle). Regrettably, neither DCPSE nor Corpuscle makes available the original audio files, because many of the recordings in LLC–1 were made surreptitiously.

The design of LLC–1 is a feature that has been of crucial importance for us in the development of LLC–2. Figure 1 presents the basic design of the earlier corpus. As can be seen in the figure, LLC–1 comprises a range of discourse contexts, which can broadly be divided into dialogue and monologue. Dialogues are either private conversations or public discussions such as interviews and panel discussions. The conversations are further divided into face-to-face conversations (recorded either surreptitiously or non-surreptitiously) and telephone conversations. Monologues are either spontaneous or prepared. The former includes improvisation, e.g. running commentaries on sports events, while the latter are planned in advance. Prepared monologue is, in turn, divided into monologue that is written to be spoken, such as political speeches, and monologue that is spoken to be written down, such as dictated letters.

The design of a corpus is its most important feature because it determines the extent to which the corpus is representative of the variability in the population from which it is sampled and balanced with regards to the proportions of the varieties of speech vis-à-vis the population. This, in turn, determines the generalisability and validity of the results obtained from the corpus data to that population. As the world's first publicly available spoken corpus, the developers of LLC–1 were faced with a challenge that no linguist had encountered before, namely, how to sample a spoken corpus in a way that respects the representativeness and balance of the corpus design. This issue was dealt with through the lens of the general goal of the Survey of English Usage, that is, 'to describe the grammatical repertoire of adult educated native speakers of British English' (Svartvik & Quirk 1980: 9) through the inclusion of different types

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

of varieties of speech and writing. Thus, the size and design of LLC–1 derive from three main considerations (summarised from Svartvik & Quirk 1980: 9–11).

- *Sample size*: 500,000 words was considered optimal.
- *Representativeness*: the varieties of speech in LLC–1 reflect the parameters affecting linguistic variation across speakers in different contexts and situations. For example, the developers found that intimacy and distance affected very strikingly the kinds of grammatical and stylistic features used in conversation, hence the distinction between participants who regarded each other as on an intimate, an equal, or on a distant footing.
- *Balance*: the speech varieties contain the relative amount of data vis-à-vis the whole corpus that is required to represent the grammatical and stylistic potential of each variety. Due to the preponderance of conversations among equals in the population, the relative amount of data was considered more useful for linguistic analysis than the statistical distribution of the varieties in the population.

As a result, LLC–1 should provide a sufficiently comprehensive basis for the study of grammatical and stylistic variation of spoken British English. Precedence was given to private face-to-face conversations among people who knew each other well. However, care was taken to add to the corpus as many other discourse contexts as possible. It is important to note that the considerations in LLC–1 were based on impressionistic reasoning rather than principled empirical investigations of linguistic variation in a pilot corpus (e.g. Biber 1993).

It is the wide range of text types in LLC–1 that most certainly has contributed to its usefulness. Linguists are not only able to study natural speech in various discourse contexts, but also to verify their intuitions about the similarities and differences in grammatical and stylistic features across the contexts (e.g. face-to-face vs telephone conversation; see Erman 1988; Aijmer 1996).[5] Early studies on LLC–1 also included comparisons with the written component of *The Survey of English Usage Corpus*, as well as other written corpora such as *The Brown Corpus* (Francis & Kučera 1964) and *The Lancaster–Oslo–Bergen Corpus* (Johansson, Leech & Goodluck 1978), which were compiled based on design criteria similar to those of LLC–1, i.e. containing samples of a sufficiently wide range of text types. However, because of its age, the use of LLC–1 seems to have decreased in spoken corpus research. With data from the 1950s to 1980s, LLC–1 is no longer a valid resource to be used as a proxy for contemporary spoken English, because change in spoken language happens fast. Moreover, there is a shortage of corpora comparable to LLC–1 to allow for diachronic comparisons of speech. DCPSE, consisting of LLC–1 and *The British Component of the International Corpus of English* (ICE–GB) from the 1990s (Nelson, Wallis & Aarts 2002), is widely used, but it does not include data from the twenty-first century. Inspired by this, we initiated the compilation of LLC–2. The design criteria of LLC–1 were instrumental in the development of this new corpus, as will become apparent in the next section.

[5] See www.ucl.ac.uk/english-usage/archives/seu-biblio.htm for a list of publications based on LLC–1.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

3 Corpus design

As already mentioned in section 2, the size and design of LLC–2 are modelled on the same principles as those of LLC–1, which means that we have by and large assumed the adequacy of the reasoning behind representativeness and balance in the earlier corpus. However, some of the design features in LLC–1 were deemed unsuitable for LLC–2 and had to be changed. In this section, we identify the main design features of LLC–2 and discuss the pros and cons of the decisions made in each case (for direct comparisons between LLC–1 and LLC–2, see section 6 below).

LLC–2 contains approximately 500,000 words of contemporary spoken British English from 2014 to 2019, stored in 100 texts of around 5,000 words each. This makes LLC–2 a *sample corpus* that consists of text samples from a range of recordings rather than a more limited number of texts of whole recordings. In a relatively small corpus such as LLC–2, an important advantage of the sample-based approach is that it provides more diversity among the speech events included in the corpus and a broader representation of linguistic phenomena (cf. Biber 1993: 252). Another advantage is that undue influences of long texts might skew the corpus results to idiosyncratic styles (cf. Sinclair 2005: 'Sampling' para. 5).

The texts in LLC–2 are either unitary texts or composite texts, meaning that they either contain material from one recording (unitary) or consist of multiple shorter recordings revolving around a similar subject matter and/or involving the same speaker(s) (composite). While most of the composite texts in LLC–2 contain two subtexts of approximately 2,500 words each (mainly private conversation), the rest can consist of as many as nine subtexts of approximately 500 words each. These short subtexts typically represent complete speech events, e.g. a science demonstration, with an obvious opening and closing. According to Sinclair (2005: 'Sampling' para. 8), this should be the 'gold standard' in spoken corpus compilation: '[s]amples of language for a corpus should wherever possible consist of entire documents or transcriptions of complete speech events, or should get as close to this target as possible'. This is because a complete text sample represents a larger context for linguistic choices, some of which only occur at the beginning, in the middle or at the end of a speech event. In order to increase the likelihood of the occurrence of such linguistic choices in longer subtexts and unitary texts in LLC–2, we selected the samples from different parts of the recordings.

Table 1 presents the complete design of LLC–2. In the table, the 100 texts (marked with 'T' followed by a unique numeric code; third column) are distributed across seven different groupings or text categories. In this article, the term 'text category' refers to the formal counterpart of discourse contexts and is used to talk about the texts in the corpus, and 'discourse context' refers to actual instances of the meaningful functioning of the utterances on the occasion of use.

Most of the text categories in table 1 are, in turn, divided into subcategories. First, face-to-face conversation takes place among equals or disparates. Equals are friends, peers in the workplace or related by descent or partnership (e.g. parent–child; husband–wife), while disparates have hierarchically unequal positions in a workplace

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

Table 1. *The complete design of LLC–2 (CMC = Computer-Mediated Communication)*

Text category	Subcategory	Text IDs	Subcat. total	Text cat. total
Face-to-face conversation	Equals	T001–T032	32	47
	Disparates	T033–T047	15	
Phone/CMC conversation	Audio	T048–T054	7	12
	Video	T055–T059	5	
Broadcast media	Discussions	T060–T063	4	12
	Interviews	T064–T071	8	
Parliamentary proceedings	Question time	T072–T073	2	4
	Debates	T074–T075	2	
Spontaneous commentary	Sports	T076–T082	7	12
	Video games	T083	1	
	Science	T084–T086	3	
	Cooking	T087	1	
Legal proceedings	Hearings	T088–T091	4	4
Prepared speech	Lectures	T092–T093	2	9
	Popular science	T094–T096	3	
	Sermons	T097–T098	2	
	Politics	T099–T100	2	
			100 texts	100 texts

or an educational institution (e.g. employer–employee; teacher–student). Second, phone/CMC conversation is carried out in such a way that the speakers can only hear each other (audio) or they can also see each other via a webcam, facilitated by the video chat provider Skype (video).[6] Third, broadcast media are taken from the radio or internet and include discussions and interviews on specific topics. Fourth, parliamentary proceedings are represented by question time and debates, delivered at the two houses of the UK Parliament. Fifth, spontaneous commentary includes commentary on sports events, video games, as well as science and cooking demonstrations. Sixth, legal proceedings are represented by hearings in the UK Supreme Court. Seventh, prepared speech includes speeches given by one person in a variety of contexts: university lectures, popular science talks, sermons and political speeches.

The text categories in LLC–2 represent different contextual constraints that may affect language use and participant behaviour. The distinctions that are commonly made in corpus linguistics are between dialogue and monologue, private and public speech, and spontaneous and prepared speech (see, for example, figure 1 about LLC–1 above). However, instead of imposing strict boundaries on the text categories, we decided to adopt the notion that whether a category is, say, a dialogue or a monologue is a matter of degree rather than either/or. For example, the categories in table 1 are ordered from the most common type of dialogue (face-to-face conversation) to the most common

[6] Note, however, that the video recordings are not available to the public; instead, they were used to facilitate the task of the transcribers.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

type of monologue (prepared speech), with the rest of the categories occupying a position somewhere in between. Thus, spontaneous commentary may be either a monologue or a dialogue, depending on whether the speech event involves one or more speakers, and whether or not the speakers are simultaneously engaged in a conversation with each other. Moreover, there is a continuum from private speech to public speech, defined as speech that was exclusive to the speakers (private) and speech that was also available to outsiders (public). While the first two categories in table 1 are mainly private and the rest are mainly public, there is some variability within the text categories themselves (e.g. some of the recordings within phone/CMC conversation are of public radio phone-ins). Finally, the first six categories are mainly spontaneous, while prepared speech is not. For more information on the text categories, see the LLC–2 user guide (Põldvere, Johansson & Paradis in press).

The sampling procedure in LLC–2 followed the principle in corpus compilation that the representation of texts in a corpus should be proportional to both their initiators (speakers) and receivers (addressees and other listeners) (e.g. Atkins, Clear & Ostler 1992; Biber 1993; Burnard 2000, 2002; Leech 2007). Therefore, the largest proportion of the texts in LLC–2 is made up of discourse contexts in which most of us are engaged on a regular basis as both initiators and receivers such as face-to-face conversation and phone/CMC conversation (59 texts in total).[7] This decision is in line with the general sampling guidelines in corpus linguistics, whereby conversational dialogue is 'the dominant component of general language both in terms of language reception and language production' (Burnard 2000: 'Design of the spoken component' para. 1), and 'private conversation merits inclusion [in a corpus] as a significant component of a representative general language corpus' (Atkins, Clear & Ostler 1992: 7). This said, the ratio between those two broad types of speech situations in LLC–2 is not proportional to the ratios that exist in the population. Had this been the case, possibly over 90 per cent of the corpus would have had to consist of private conversation (cf. Svartvik & Quirk 1980: 11; Biber 1993: 247). However, such a corpus would miss out on the linguistic variation present in discourse contexts to which many of us are exposed as receivers, despite the fact that we rarely, if at all, participate in them as initiators. Therefore, a considerable, yet smaller, proportion of LLC–2 consists of discourse contexts that are in the public domain (41 texts in total). The best-represented text categories in this group are broadcast media and spontaneous commentary with 12 texts in each category.

The principle of proportional sampling relative to initiators and receivers was also considered in the inclusion of individual texts in the text categories. Thus, we took into account, albeit impressionistically, factors such as popularity and impact. For example, broadcast media contain, among other things, podcasts that, according to various rankings, have a large listener base in the UK (e.g. The Adam Buxton Podcast; but see section 4.1.2 below for issues with copyright). Similarly, legal proceedings contain

[7] Note that radio phone-ins within phone/CMC conversation do not strictly meet this criterion. However, some of the specific phone-ins in LLC–2 include non-professionals who at least have the possibility to call a radio station if they wish, which means that they may act as initiators in such contexts on a regular basis.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

recordings of hearings in the UK Supreme Court, the final court of appeal in the UK, where cases of the greatest public or constitutional importance are discussed, rather than one of the lower-level courts such as local county courts.

One possible shortcoming of the proportions in LLC–2 is that some of the text (sub) categories are not represented well enough. As can be seen in table 1, three of the text categories in the corpus contain fewer than 10 texts (parliamentary proceedings, legal proceedings, prepared speech), which may not be enough for representative studies of certain types of variation. Additional problems arise if users wish to carry out investigations of the subcategories, in which case they may have available to them only one text (e.g. cooking demonstrations within spontaneous commentary). This problem is, of course, due to the presence of an upper limit on texts in LLC–2 (i.e. 100), which is a predetermined feature of the design of the corpus that constrains the inclusion of more texts in the (sub)categories. In order to mitigate the problem somewhat, we have included as many subtexts as possible within those subcategories, so that if a subcategory contains very few texts, these texts are made up of several recordings in order to increase variability (e.g. five different cooking sessions in one text of cooking demonstrations). This said, users are advised to carry out investigations of the text categories rather than their subcategories in order to ensure better representation of linguistic features or, alternatively, to combine the subcategories with data from other sources to build a specialised dataset (e.g. a DIY corpus of cooking demonstrations).

All in all, then, the approach to the notions of representativeness and balance in LLC–2 is to view them as continua rather than all-or-nothing (see Leech 2007, but also Põldvere 2019). Following Leech (2007: 144), we have sought to 'define realistically attainable positions on these scales', rather than setting unrealistic goals or abandoning the notions altogether. As a result, we have a corpus that, like LLC–1, provides a sufficiently representative account of linguistic variation in contemporary spoken British English with evident regard for the distribution of the text categories. In section 6 below, we will examine another important notion of corpus design that needed to be considered, namely, comparability.

4 Data collection, ethics and speakers

This section describes the decisions made in the collection of data for LLC–2, including the ethical and legal considerations that we had to take into account (section 4.1). While section 4.1.1 describes the procedure for collecting respondent recordings, section 4.1.2 does the same for web-based recordings. In section 4.2, we summarise the main demographic information about the speakers who appear in the corpus.

4.1 Recordings

The main challenge of collecting data for LLC–2 was deciding what types of speakers should be targeted. While the LLC–1 corpus design gave us a good idea of the types of text categories that should be included and to what extent, there was very little

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

guidance as to the demographic information about the speakers. One of the few criteria in LLC–1 was that the speakers should be educated adult speakers of British English. Accordingly, the lower age limit in LLC–2 was set to 18 (but see section 4.2 below for exceptions) and the recordings were collected in locations in the UK where educated adults could be recruited (e.g. universities). Nevertheless, much of the data collection for LLC–2 was opportunistic; instead of following a strict sampling frame for speaker demographic information, we accepted recordings from anyone who was interested in contributing to the project as long as they met the selection criteria above. It was only at a later stage that imbalances in the data recorded were reduced to the extent possible. The two-part data collection procedure adopted in LLC–2 was very useful for this purpose. Specifically, the recordings were collected in one of two ways: (i) respondent recordings, where we asked people to make recordings specifically for the corpus,[8] and (ii) web-based recordings, where we collected the recordings from different sources on the internet. Each recording type has its own possibilities and limitations, and they make different allowances for the choice of speakers. In what follows, we explain the procedure for each type, starting with respondent recordings.

4.1.1 Respondent recordings

Respondent recordings in LLC–2 cover roughly three text categories. They include all the face-to-face conversations, most of the phone/CMC conversations and all the university lectures within prepared speech. The collection of the recordings was mainly carried out at University College London (UCL). The reason for this was that UCL was also the main site of recording for conversations in LLC–1 (i.e. face-to-face conversation and telephone conversation). In addition, a few recordings were made at another university in England, Lancaster University, and at Lund University in Sweden. The recordings in Sweden were university lectures given by native speakers of British English. As a result, most of the speakers in respondent recordings seem to have been from England, with a high concentration of people from London; however, since the areas in which we recorded attract people from all over the UK, we also managed to record speakers from other constituent parts (Scotland, Wales, Northern Ireland).

As previously mentioned, the collection of respondent recordings in LLC–2 was largely opportunistic. We were well aware of the challenge of recruiting people and persuading them to record a minimum of 30 minutes of conversation or a lecture without any compensation. The recruitment of people was advertised in several ways: (i) personal networking; (ii) physical posters and leaflets on and near university campuses; and (iii) posts on online mailing lists and social media pages. It is difficult to say which approach yielded the best results; rather, it was the combination of efforts that helped us reach our goal. While advertisements on campus ensured that the speakers in LLC–2 share many of the same characteristics with those in LLC–1, online advertising and physical networking helped us reach out to people from other areas and

[8] Note that we use the term 'respondents' to refer to people who were responsible for making the recordings, and 'speakers' to refer to all the people in the recordings.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

backgrounds. This allowed us to diversify the pool of speakers in the corpus and thereby obtain a more representative sample.

Another way to ensure a diverse pool of speakers in LLC–2 was to complement the opportunistic data collection procedure with selective targeting of certain demographic groups. This was done at a later stage of the project when it became clear that certain groups of speakers were more willing than others to respond to our advertisements. This was particularly the case with young university students, who were then asked to record their conversations, not with their peers, but, for instance, with their parents and grandparents. A similar approach was adopted in the compilation of *The Spoken British National Corpus 2014* (BNC2014; Love *et al.* 2017), but not in *The Spoken British National Corpus 1994* (BNC1994; BNC Consortium 2007), which, instead, followed a sampling frame of selection criteria such as the age, gender and socio-economic status of the speakers. However, not even a large corpus such as the BNC1994 can meet all the target proportions, which seems to have triggered a more opportunistic approach: 'to collect as much data as possible and to accept the consequent imbalances in the corpus across the demographic categories' (Love *et al.* 2017: 326).

Furthermore, an important issue to consider in spoken corpus projects is the choice of recording equipment. In LLC–2, most of the face-to-face conversations and university lectures were recorded with high-quality equipment, including a digital voice recorder (Zoom H4n Handy Recorder), and, if necessary, an external microphone and a small action video camera. The decision to use high-quality recording equipment instead of, say, smartphones (see Love *et al.* 2017 for the BNC2014) was mainly due to our goal to make LLC–2 suitable for prosodic analyses. Despite advances in the recording quality of smartphones, digital voice recorders are still better suited for recordings in noisy environments as they filter out background noise. The use of our own equipment also had an unexpected benefit, because through the meetings with the respondents to distribute and collect the equipment, we gained unique insight into the recording situations and the speakers, which turned out to be helpful during transcription later on.

Nevertheless, at a later stage of the project the respondent recordings were made either with smartphones or the respondents' own digital voice recorders. The reason for this was that the data collection continued beyond our trips to the UK. The need to distribute and collect the recording equipment is, of course, an important disadvantage of using one's own equipment. This is different from smartphones, which considerably reduce the cost and effort of recording private conversations (Love *et al.* 2017: 329). However, in our case both approaches were needed to reach the goal. The phone/CMC conversations in LLC–2 were also recorded at a distance, because the software needed to record mobile phone calls and Skype video calls could easily be downloaded from the internet (see also Diemer, Brunner & Schmidt 2016).

The ethical and legal considerations of collecting respondent recordings in LLC–2 were addressed by seeking written permission from all the speakers prior to recording. This was different from LLC–1 where many of the conversations were recorded without prior knowledge of the speakers (see section 2 above). The permissions in

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

LLC–2 were requested via a consent form, which the respondents distributed among the speakers in the recordings. In addition to information about the project and the recording task, the consent form contained a statement of the transfer of rights to the recording to the corpus developers, giving us the permission to transcribe the recordings and make both the transcriptions and the original recordings publicly available for non-commercial use. The consent form was developed together with a legal expert at Lund University and, if required by the institution in which the recordings were made, it was approved by the appropriate ethics committee.

4.1.2 Web-based recordings

Web-based recordings cover the rest of the text categories in LLC–2: all the broadcast media, parliamentary proceedings, spontaneous commentary, legal proceedings and the remaining recordings within prepared speech (i.e. popular science talks, sermons, political speeches, but not university lectures). In addition, they cover a few radio phone-ins within phone/CMC conversation. As previously mentioned, web-based recordings in LLC–2 were obtained from various sources on the internet. This allowed for a more targeted approach to data collection, because the wealth of data found online allowed us to seek out speakers in demographic groups with the largest imbalances (e.g. older speakers). Efforts were also made to achieve a more balanced representation of geographical region by seeking out speakers from different constituent parts of the UK. However, this information is not part of the speaker metadata for web-based recordings and, therefore, it is not readily available for analysis either (see section 4.2 below).

The main challenge of collecting web-based recordings in LLC–2 was to secure copyright permissions for the recordings that we had chosen. Specifically, we requested permission to include in the corpus, and thus distribute as part of it, both the original recordings and the transcriptions prepared by us. However, most of our requests were either rejected or ignored. This was often the case with major media organisations, which would have been the best candidates for achieving a representative corpus in terms of popularity and impact (e.g. BBC, Ted Talks; see section 3 above). Surprisingly, the main reason for the rejections was that the media organisations only licensed material for commercial use, not for non-commercial use, which would have applied to our needs. Such setbacks led us to seek out other, less hierarchical organisations where decisions about copyright and licensing were made more promptly and with fewer restrictions. The permissions were granted either in the form of a written statement or we were asked to sign a licence agreement which outlined the terms and conditions of use of the material. The recordings obtained from these sources turned out to be equally suitable for inclusion in LLC–2.

4.2 Speaker characteristics

This subsection presents the main demographic information about the speakers in LLC–2. This information allows us to examine the outcome of the two-part data collection

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

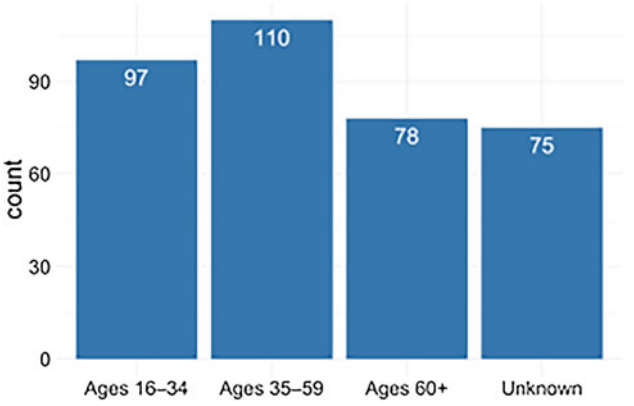

Figure 2. The distribution of speakers across four age groups in LLC–2

procedure described above. In respondent recordings, the demographic information was obtained by asking all the speakers in the recordings to fill in a questionnaire, which contained questions about the speakers' age, gender, occupation, education, (foreign) language use, place(s) of residence and accent. Due to issues of reliability with extracting personal information from the internet, the web-based recordings only include the speakers' age, gender and occupation. Below, we present the most important demographic information available for both types of recordings: age and gender.

Figure 2 illustrates the distribution of the 360 speakers in LLC–2 across four age groups: 16–34, 35–59, 60+ and Unknown.[9] The age groups are specific enough to illustrate generational differences and broad enough to prevent data sparseness. As can be seen in figure 2, the three age groups for which we have data are distributed relatively evenly in LLC–2. The age group 35–59 has 110 speakers and the age groups 16–34 and 60+ have 97 and 78 speakers, respectively. Thus, the combination of opportunistic data collection with some selective targeting at the end of the data collection gave us a sufficiently balanced sample. This said, there is a large number of speakers, 75, in mainly web-based recordings whose ages could not be reliably determined. This issue could have been alleviated by seeking out public figures with greater internet presence; however, such recordings may have been more difficult to obtain due to copyright issues (see section 4.1.2 above).

In contrast to age, the demographic category of gender shows clear differences. Specifically, there are many more male speakers, 221, in LLC–2, than female speakers, 139.[10] On closer inspection, this seems to be mainly due to the predominance of male

[9] Note that, despite the lower age limit of 18, two of the speakers in LLC–2 are 16 and 17 years old. Considering that the recordings in which they appeared also included their parents as their legal guardians, and that written consent was obtained from all speakers, we decided to keep the recordings anyway.

[10] In the questionnaires, the speakers were also given the option to choose a gender other than male or female; however, this option was not chosen by anybody.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

speakers in certain professional contexts such as parliamentary interactions and court hearings. Despite making up only a small proportion of LLC–2, these text categories contribute a considerable number of speakers. Thus, different distributions may exist for different text categories, or when the distribution of word tokens rather than speakers is considered (see Põldvere, Johansson & Paradis in press).

5 Transcription, markup and annotation

The challenges described in relation to data collection in section 4 are minor in comparison to the time, effort and knowledge required to turn the original audio recordings into written form, that is, to transcribe them. As many scholars have pointed out, transcription is not a neutral process, but rather it is fundamentally selective, interpretive and reflective of the corpus developers' own research interests (e.g. Ochs 1979; Du Bois 1991; Edwards 1995; Bednarek 2020). Added to this, corpus developers need to consider the methodological and technological challenges of transcription such as reliability, searchability and interoperability (Schmidt 2016). In the compilation of LLC–2, we were guided by two main principles: (i) the corpus needs to be useful for a wide range of linguistic studies including prosody, and (ii) the transcriptions need to be formatted in such a way that they are compatible with widely used corpus linguistic and text processing tools for easy retrieval and analysis.

This section describes how the principles were realised in the transcription, markup and annotation of the corpus. Section 5.1 starts by presenting the main features of the transcription and markup scheme used in LLC–2, followed by an overview of how the original transcriptions were converted into a fully canonical XML format in section 5.2. Finally, section 5.3 provides information on further annotations.

5.1 Transcription and markup scheme

The transcription and markup scheme of LLC–2 is based on enhanced orthographic transcription (Crowdy 1994), which involves the transcription of orthographic words enhanced by markups of basic spoken features such as (filled) pauses, overlaps, non-verbal vocalisations, events and so on (see Põldvere, Johansson & Paradis in press for the full scheme). Enhanced orthographic transcription was chosen because it meets the first principle mentioned above: it is applicable for a wide range of linguistic studies such as lexicology, morphology, syntax and discourse analysis (but see section 5.3 below for prosody). The precise nature of a transcription and markup scheme hinges on the key requirement for the scheme to minimise 'the level of transcriber inference that is needed – that is, the number of decisions that a transcriber must make about potentially ambiguous speech phenomena' (Love *et al.* 2017: 334). After a close review of several schemes, we decided to use *The International Corpus of English* (ICE) as a model, which already had formed the basis of more than twenty corpora of English worldwide including ICE–GB. The ICE conventions (Nelson 2002) have been shown to provide rich linguistic research material while being simple enough to

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

encourage reuse by other corpus developers. This said, some simplifications were needed to further reduce transcriber inference.

A central feature of the transcriptions in LLC–2 is that they are segmented by speaker turns rather than, as in ICE, by orthographic sentences. Defined as verbal units bounded by the talk of another speaker (Ochs 1979: 69), speaker turns are the basic structure of speech transcription (Atkins, Clear & Ostler 1992: 11) and, as such, relatively easy to identify. This is different from orthographic sentences, which in spoken language do not always correspond to well-formed grammatical sentences in writing, but may appear as speech fragments. Each speaker turn in LLC–2 is preceded by a unique speaker ID and a timestamp; in very long contributions by one speaker, the timestamps were inserted every minute. This so-called audio-to-text alignment is an important innovative feature of the corpus that specifies where in the recording the turn begins (see Põldvere, Frid, Johansson & Paradis 2021 for details).

Furthermore, there are a number of features in ICE that, in our experience, pose problems for reliable encoding of spoken texts and because of that they were not included in LLC–2. One such feature is direct reported speech. It became clear during the transcription of LLC–2 that direct reported speech cannot be reliably captured without access to the original source. Consider (1):[11]

(1) <S044> all I literally need to know is like do you wanna be friends okay so he's answered that yes okay part one's out the way two do you even wanna be the type of friends that I'm talking about right cause there's obviously different types of friendship right apparently the answer to that is yes three are you committed to actually doing what it takes to do that to achieve that yes or no

In (1), speaker <S044> reads a text message from an ex-boyfriend out loud to her current partner. While doing so, she switches between the actual text message and her comments on it. Since we do not have access to the original text message, it is not possible for us to know what is what. For example, it is unclear whether the text message contains the phrase *to do that* or *to achieve that*, or both. Therefore, our encoding of the direct reported speech in (1) may have differed considerably from the intuitions of the end users. Other similarly problematic features that were not included in LLC–2 were punctuation, which according to Schiffrin (1994: 25) may be too interpretive and/or misleading, and different lengths of pauses, the strict measurement of which is highly time-consuming (Thompson 2005: 'Transcription' para. 6; but see section 5.3 below).

A feature of spoken language for which we decided not to make any compromises was overlapping speech. In contrast to direct reported speech, overlaps (and gaps) are fundamental components of dialogic spoken interaction. As a result, the transcriptions in LLC–2, like in ICE, contain detailed markings of the start and end points of overlaps in spite of the fact that sophisticated markup of overlaps requires substantially more time and effort than a simplified encoding (Atkins, Clear & Ostler 1992: 12), as

[11] The speaker IDs are the same as in the corpus.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

in when the turn is marked up according to whether or not it overlaps with the immediately preceding turn (e.g. BNC2014). Furthermore, the sophisticated markup of overlaps in LLC–2 allowed us to preserve the right sequencing of speaker turns in the conversation. Consider (2). The overlaps in the example are represented by square brackets with multiple overlaps being matched by numbers.[12]

(2) <S130> we're doing the same things as last year then [1][really]
<S129> [1][same] [2][yeah so] I mean the same the making the induction packs
<S130> [2][yeah cause]

The transcription in (2) contains two sets of overlaps, first between *really* and *same* and then between *yeah so* and *yeah cause* (an incomplete turn). This is to accurately represent the sequential unfolding of the turns where <S130>'s backchannel *yeah* is a response to what <S129> said earlier (*same*).

The second principle, that is, to format the transcriptions in such a way that they are compatible with various tools, was met by encoding the transcriptions in the standardised markup language 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 the angle brackets is the actual corpus text. This allows for easy extraction of relevant linguistic information and compatibility with well-known corpus linguistic and text processing tools such as *AntConc* (Anthony 2020) and *Wordsmith Tools* (Scott 2020). Moreover, the integration of the XML standard already in the transcription stage rather than in the postprocessing stage allowed us to keep track of the word count of the actual corpus text in a reliable way (see section 3 above). Now, XML is not particularly transcriber-friendly, because inserting XML tags is highly time-consuming for a human transcriber. To mitigate this problem, we opted for a two-step procedure whereby, in the first step, the transcribers manually inserted simplified XML tags into the transcriptions, and, in the second step, the transcriptions were converted into a fully canonical XML format in a semi-automatic fashion. It is the outcome of the second step that constitutes the final version of LLC–2. The rest of this subsection provides a brief overview of the procedure in the first step, and section 5.2 below contains a more detailed description of the standard in the second step.

The first step of transcribing and marking up LLC–2 was carried out in the transcription software *InqScribe* (2005–20). The software allows for quick insertions of pre-defined snippets, which in our case included both the markups of the spoken features and the timestamps mentioned above. However, achieving a reliable transcription was a major concern in this step. Despite adopting a transcription and markup scheme that considerably reduced transcriber inference, several other safeguards needed to be put in place. For example, the most complex transcriptions in LLC–2 (e.g. face-to-face conversation) were made by two people, who followed detailed instructions (see

[12] Note that the example does not include timestamps in order to facilitate the task of the reader of this article.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

Põldvere, Johansson & Paradis in press) and who underwent rigorous training. The transcribers worked independently on different transcriptions, which were then checked by the other transcriber. Disagreements were discussed and resolved together. Moreover, many of the other transcriptions were facilitated by access to already existing online transcriptions (e.g. Hansard for parliamentary proceedings), which in addition provided a useful template against which to check our own intuitions and judgments. Finally, the availability of the original audio recordings (see section 7 below) makes it possible for users to carry out their own checks before analysing the corpus data.

5.2 XML conversion

After the first step, the transcriptions were converted into a fully canonical XML format in the second step in order to improve the searchability and interoperability of LLC–2, and to produce the final XML version of the corpus. The conversion was carried out with the help of various in-house scripts, which were specifically developed for this purpose. The conversion of the transcriptions to a single consistent format and validation of its structure was greatly facilitated by the XML standard adopted in the first step. It allowed for an unambiguous mapping of the preliminary tags into fully XML-compliant tags. This second step of the procedure was also used to make additional changes to the transcriptions, which were not possible prior to conversion to a structured document (see below).

The structure of the final XML version follows closely the recommendations in Hardie's (2014) 'Modest XML for corpora', designed to provide an alternative to well-known standards of corpus encoding, most notably, the Text Encoding Initiative (TEI) and the Corpus Encoding Standard (CES). According to Hardie, these standards are unnecessarily complex for most corpus compilation projects. For example, the minimal TEI header is 'a very large, complex block of markup', the insertion of which requires considerable effort from the corpus developers (Hardie 2014: 77). The modest XML system developed by Hardie consolidates this complexity by offering reasonably standard and easy-to-understand ways of inserting XML tags. Moreover, the system is meant to be viewed as an open-ended set of suggestions rather than a standard, thus leaving ample room for flexibility. This is particularly important in the light of the fact that only a few of Hardie's suggestions are based on spoken features. The approach adopted in LLC–2 was to adhere to Hardie's system as closely as possible, while at the same time being mindful of the requirements of our own transcription and markup scheme.

For example, in order to provide users with easy access to information about the speech event and the demographic details of the speakers, we adopted a much simpler way of inserting headers than in the TEI standard. The headers in LLC–2 are less hierarchical, consisting of a set of tags that each correspond to one piece of metadata information (see also the BNC2014 for the same approach). In order to stay true to our transcription and markup scheme, we also made changes in the body of the

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

transcriptions. The most important of these was the introduction of the <turn> element to reflect the segmentation of speech in LLC–2 by speaker turns. Following Hardie's (2014) recommendations, the element contains the attributes *n* and *who* (the sequential position of the turn and speaker ID, respectively), but we also added to it the *timestamp* attribute. In cases where Hardie's recommendations did not cover the spoken features in LLC–2, other resources were consulted. This was, for example, the case with overlaps. Overlaps are notoriously difficult to represent in XML because different overlaps tend to intrude on each other's regions, which is unacceptable in XML. Thus, following Weisser (2017), we used the single element <overlap> containing the attributes *pos* and *n*, which indicate whether the tag marks the *start* or *end* of the overlap (*pos*) and which regions occur together (*n*). Example (3) illustrates the use of the <turn> (in bold) and <overlap> (underlined) elements in the final version of LLC–2. The example is the same as in (2) above.[13]

(3) 1 **<turn n="17" timestamp="00:01:29.08" who="S130">**we're doing the same things as last year then <overlap pos="start" n="15"/>really<overlap pos="end" n="15"/>**</turn>**
2 **<turn n="19" timestamp="00:01:35.00" who="S129"**><overlap pos="start" n="15"/>same<overlap pos="end" n="15"/> <overlap pos="start" n="16"/>yeah so<overlap pos="end" n="16"/> I mean the same the making the induction packs**</turn>**
3 **<turn n="20" timestamp="00:01:35.18" who="S130"**><overlap pos="start" n="16"/>yeah cause<overlap pos="end" n="16"/>**</turn>**

5.3 Further annotations

In addition to XML annotation, the transcriptions in LLC–2 were also annotated for parts-of-speech (POS) and lemma information. Owing to the fact that it serves as a basis for a wide range of linguistic studies, POS tagging is the most widely used annotation type in corpus linguistics (Gries & Berez 2017: 383). We used the CLAWS tagger (Garside 1987) to assign each word in the transcriptions a POS tag with the highest probability. The tagset used in LLC–2, known as C7, contains a total of 140 tags.[14] The POS and lemma information in the corpus release are given as attributes in the XML element <w> assigned to each word token. The reason for choosing CLAWS was because the linguistic resources in the tagger have also been trained on spoken sources, and thereby CLAWS has been shown to achieve a high degree of accuracy for spoken texts. For example, the tagger's error rate during the compilation of the BNC2014 was only 2.5 per cent, which is comparable to that of written texts (*The British National Corpus 2014* 2018: 66). It is our hope that the POS and lemma information in LLC–2 will form the basis of more complex annotations of the corpus, for example, syntactic parsing, and semantic and pragmatic annotation.

[13] The numbers in the *n* attributes do not follow each other sequentially because, in the original transcription, there is intervening text between the speaker turns shown here.
[14] See http://ucrel.lancs.ac.uk/claws7tags.html

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

Perhaps the most important innovation in LLC–2 is that annotation was applied, not only to the transcriptions, but also to the corresponding audio files. More precisely, the speech signal of the audio files was annotated in order to facilitate the anonymisation of personal information in the recordings and to make them publicly available (see section 7 below for details on the release). The technique adopted is based on a Praat script written and developed by Hirst (2013). The script involves manual identification of all personal information in the audio files and subsequent automatic replacement of these portions of the speech signal with a *hum* sound that makes the lexical information incomprehensible, while at the same time retaining the pitch and intensity envelope of the original (see Põldvere, Frid, Johansson & Paradis 2021 for details). In this way, the anonymisation of the audio files builds on that of the transcriptions, where the personal information was changed by retaining the word class and the number of syllables of the original. Thus, the anonymised audio files make possible prosodic analyses of LLC–2. The release of the audio files, in general, makes it possible for users to annotate the corpus for linguistic features not already captured by the orthographic transcriptions (e.g. different lengths of pauses).

6 Comparisons with *The London–Lund Corpus 1*

Section 2 was concerned with representativeness and balance. This section focuses on a third notion: comparability. This follows from the fact that, in addition to being a corpus of contemporary spoken British English, LLC–2 was modelled on the same principles as those of LLC–1 in order to facilitate principled diachronic investigations of spoken language across time. The following subsections break down the main similarities and differences between the London–Lund Corpora (i.e. LLC–1 and LLC–2) in terms of their designs and speakers (section 6.1), and transcription and markup conventions (section 6.2). In addition to LLC–1, LLC–2 is comparable to other corpora of spoken British English (e.g. DCPSE). Due to major differences in the sampling frame, LLC–2 is less suited for comparisons with national corpora (e.g. BNC2014) or corpora designed to represent a specific discourse context. However, data from these corpora can be combined with one of the many text categories in LLC–2.

6.1 Corpus designs and speakers

Broadly speaking, it can be argued that, design-wise, the London–Lund Corpora differ from each other in terms of one parameter only, namely, the parameter of time (Leech 2007); there is a difference of approximately fifty years between the two corpora. However, there are minor differences within the text categories. The differences are primarily related to the conflict between the notions of comparability and representativeness: '[a]s one nears to perfection in comparability, one meets with distortion in terms of representativeness, and vice versa' (Leech 2007: 142). One text category in which we encountered this problem was so-called distanced conversations, that is, landline telephone conversations in LLC–1 and mobile phone/CMC

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

Table 2. *The comparison of the number of texts in the London–Lund Corpora*

Text category	Subcategory	LLC–1	LLC–2
Face-to-face conversation	Equals	41	32
	Disparates	10.5	15
Distanced conversation	NA	11	12
Broadcast media	NA	12.5	12
Parliamentary proceedings	NA	3	4
Spontaneous commentary	NA	12	12
Legal proceedings	NA	3	4
Prepared speech	NA	7	9
		100 texts	100 texts

conversations in LLC–2. Using landline telephone conversations in LLC–2 to ensure comparability with LLC–1 would have come at the expense of LLC–2 as representative of the communication channels more commonly used in the twenty-first century. According to a survey by the communications regulator Ofcom, landline phone calls have become increasingly obsolete in the past few decades, while mobile data use and internet services have soared (Sweney 2019). Thus, in the case of distanced conversation, representativeness was ranked higher in priority than having truly comparable datasets. At other times, the opposite was true: we did not include any discourse contexts in LLC–2 that did not exist in LLC–1 because they would not have created opportunities for comparison between the corpora (e.g. service encounters). Adjustments of this kind were necessary in order to maintain the integrity of LLC–2 as a corpus in its own right, but still achieve a sufficiently high degree of comparability with LLC–1.

Efforts were also made to achieve a high degree of comparability in relation to balance. Table 2 presents a comparison of the number of texts in the seven text categories in the London–Lund Corpora. Due to intra-categorial differences between the corpora, only the subcategories of face-to-face conversation are included. The LLC–1 figures have primarily been taken from Svartvik *et al.* (1982: 20). However, their figures were not based on the whole corpus or the same type of categorisation. Therefore, the figures in table 2 are partly based on our own interpretation of what text belongs to what text category. It should also be noted that distanced conversation in LLC–1 includes not only landline phone calls, but also one Ansaphone recording (a brand name for an answering machine mainly in the UK) and one radio phone-in.

As can be seen in table 2, the texts are by and large distributed in the same way across the two corpora. The only noticeable difference is within face-to-face conversation (51.5 in LLC–1 vs 47 in LLC–2).[15] The lower number of texts of face-to-face conversation in

[15] Two of the figures in table 2 are given as decimals because the texts corresponding to them are subtexts and not complete texts.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

LLC–2 was due to the practical difficulties and cost implications of collecting private recordings within the short time frame of the project (compare four decades in LLC–1 with only five years in LLC–2). Moreover, the main reason for collecting a larger proportion of conversations among disparates in LLC–2 (15 vs 10.5) was to even out differences with conversations among equals in order to ensure more robust comparisons between the two types of face-to-face conversation.

The speakers in the London–Lund Corpora are educated adult speakers of British English, primarily from England, and many of the speakers in face-to-face and distanced conversation are associated with the same university, University College London, through either work or study. However, it is difficult to make more detailed comparisons between LLC–1 and LLC–2 because of the lack of relevant metadata in the earlier corpus. For example, there are 360 speakers in LLC–2, but it is unclear how many speakers there are in LLC–1. At first sight, the list of speakers in Greenbaum & Svartvik (1990) provides an approximate figure (around 500–600 speakers). However, on closer inspection of the audio files, it was revealed that many of the speakers appear in more than one text, which is not reported in the corpus documentation. Instead, we know which speakers appear in more than one subtext within a text, but this is not enough for us to determine the exact number of unique speakers in LLC–1. Moreover, the metadata in LLC–1 are limited to the age, gender and occupation of the speakers, and also this information is problematic. For example, age in the corpus 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 (e.g. a vegetarian) rather than what they do for a living. This is different in LLC–2, which, as mentioned in section 4.2 above, includes the above categories as well as information about the speakers' level of education, (foreign) language use, place(s) of residence and accent. The paucity of metadata in LLC–1 is an unfortunate difference between the London–Lund Corpora, which users need to keep in mind when using the corpora in diachronic research.

All in all, though, there is considerable overlap between the sampling frames of LLC–1 and LLC–2. Therefore, any significant linguistic differences between the corpora may be attributed to temporal differences between the two time periods of English rather than variability within the corpora themselves. This said, users are reminded to view comparability in the same way as its related notions of representativeness and balance – as a continuum rather than all-or-nothing – and thereby critically examine the extent to which results obtained from the London–Lund Corpora are truly comparable.

6.2 Transcription and markup

Despite considerable overlap in the sampling frames of the London–Lund Corpora, there are important differences in the way that the recordings in the corpora have been transcribed and marked up for linguistic features. First, there are differences in the transcription and markup conventions used. While the transcriptions in LLC–2 are orthographic and marked up for basic spoken features, LLC–1 contains detailed prosodic transcriptions. The lack of prosodic annotation in LLC–2 is compensated for

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

by access to the original audio recordings (see section 7 below), which allows users to carry out their own annotations using any of the existing phonetics software (e.g. *Praat*; see Boersma 2001).

Second, the format in which we captured and distributed the transcriptions in LLC–2 is better suited for contemporary corpus linguistic investigations than that of LLC–1. As mentioned in section 5 above, the transcriptions in LLC–2 are encoded in XML, making them compatible with modern corpus linguistic and text processing tools. The LLC–1 transcriptions are incompatible with such tools due to the lack of a uniform structure for representing the different spoken features.[16] For example, overlaps in LLC–1 may be represented either by asterisks or by plus signs (*yes* or +yes+), with other features being represented by yet other symbols (e.g. round brackets for different kinds of contextual comments: (laughs)). This incompatibility with modern tools is the main reason why we decided not to follow the transcription and markup scheme of LLC–1, and instead opted for a fully XML-compliant scheme in LLC–2.

7 Access and concluding remarks

Access to the transcriptions and audio material of LLC–2 follows state-of-the-art practices of corpus archiving, distribution and preservation (cf. Wynne 2005a) in that the corpus has been released for free from trusted institutional repositories for use in two different ways: (i) the XML transcriptions and the corresponding uncompressed WAV files can be downloaded in full from the corpus server at Lund University Humanities Lab (itself a node in Swe-Clarin), and (ii) basic corpus linguistic techniques such as query searches, and concordance and collocation analyses, can be implemented on LLC–2 in the corpus management and analysis system Corpuscle within CLARINO, which also facilitates an audio playback of the segments of interest. In this way, Corpuscle is home to both LLC–1 and LLC–2, thus providing further opportunities for smooth comparisons between the two corpora. Both the downloadable and the online release of LLC–2 contain metadata information about the texts and speakers, as well as a guide to using LLC–2 (Põldvere, Johansson & Paradis in press), which provides a more detailed account of the methodological decisions made in the compilation of the corpus than could be given here. Furthermore, both repositories from which LLC–2 has been released are integrated with CLARIN and therefore provide secure and long-term storage of the corpus data. The combination of a downloadable corpus and a web-based interface allows for both close reading of the complete texts and close listening, as well as more quantitative searches of the data using either Corpuscle or any of the freely available corpus linguistic and text processing tools.

To conclude, the aim of this article has been to describe and critically examine the challenging task of compiling LLC–2 from start to end by explaining and problematising the methodological decisions made in each stage and highlighting the

[16] An XML version of LLC–1 exists, but, to the best of our knowledge, it is not available to the broader research community.

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

innovations in the new corpus. The most important feature of LLC–2 is its suitability for both synchronic investigations of contemporary spoken British English across different discourse contexts and groups of speakers, as well as for principled diachronic research on spoken language over the past half a century. An important innovation of the corpus is the release of the orthographic transcriptions together with the original audio recordings in order to allow users to further enhance the value of LLC–2 relative to their own research interests (e.g. prosody, turn-taking; see Põldvere & Paradis 2019, 2020; Põldvere, Johansson & Paradis 2021). These new and exciting research opportunities will certainly be worth the long and bumpy road of compiling LLC–2 as described in this article.

Authors' addresses:
Centre for Languages and Literature
Lund University
Box 201
221 00 Lund
Sweden
nele.poldvere@englund.lu.se
victoria.johansson@ling.lu.se
carita.paradis@englund.lu.se

References

Aarts, Bas, Joanne Close & Sean Wallis. 2013. Choices over time: Methodological issues in investigating current change. In Bas Aarts, Joanne Close, Geoffrey Leech & Sean Wallis (eds.), *The verb phrase in English: Investigating recent language change with corpora*, 14–45. Cambridge: Cambridge University Press.

Aijmer, Karin. 1996. *Conversational routines in English: Convention and creativity*. London: Routledge.

Anthony, Laurence. 2020. *AntConc*, version 3.5.9. Tokyo: Waseda University. www.laurenceanthony.net/software (accessed 9 April 2021).

Atkins, Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7(1), 1–16.

Bednarek, Monika. 2020. The *Sydney Corpus of Television Dialogue*: Designing and building a corpus of dialogue from US TV series. *Corpora* 15(1), 107–19.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4), 243–57.

BNC Consortium. 2007. *The British National Corpus*, version 3 (BNC XML Edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. www.natcorp.ox.ac.uk/ (accessed 1 November 2020).

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341–5.

The British National Corpus 2014: User manual and reference guide, version 1.1. 2018. http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf (accessed 1 November 2020).

Burnard, Lou (ed.). 2000. *Reference guide for the British National Corpus (world edition)*. www.natcorp.ox.ac.uk/archive/worldURG/index.xml (accessed 1 November 2020).

https://doi.org/10.1017/S1360674321000186 Published online by Cambridge University Press

Burnard, Lou. 2002. Where did we go wrong? A retrospective look at the British National Corpus. In Bernhard Kettemann & Georg Marko (eds.), *Teaching and learning by doing corpus analysis*, 51–71. Amsterdam: Rodopi.
Clark, Herbert H. 1996. *Using language*. Cambridge: Cambridge University Press.
Crowdy, Steve. 1994. Spoken corpus transcription. *Literary and Linguistic Computing* 9(1), 25–8.
Diemer, Stefan, Marie-Louise Brunner & Selina Schmidt. 2016. Compiling computer-mediated spoken language corpora: Key issues and recommendations. *International Journal of Corpus Linguistics* 21(3), 348–71.
Du Bois, John W. 1991. Transcription design principles for spoken discourse research. *Pragmatics* 1(1), 71–106.
Du Bois, John W. 2014. Towards a dialogic syntax. *Cognitive Linguistics* 25(3), 359–410.
Edwards, Jane A. 1995. Principles and alternative systems in the transcription, coding and mark-up of spoken discourse. In Geoffrey Leech, Greg Myers & Jenny Thomas (eds.), *Spoken English on computer: Transcription, mark-up and application*, 19–34. London: Routledge.
Erman, Britt. 1988. *You know* in face-to-face and telephone conversation. In Inger Henrysson & Gunnar Persson (eds.), *English today: Papers read at the English Studies Conference in Umeå*, 14–21. Umeå: Department of English, University of Umeå.
Francis, Nelson W. & Henry Kučera. 1964. *Manual of information to accompany A Standard Corpus of Present-Day Edited American English, for use with digital computers*. Providence, RI: Department of Linguistics, Brown University. http://icame.uib.no/brown/bcm.html (accessed 1 November 2020).
Garrod, Simon & Martin J. Pickering. 2004. Why is conversation so easy? *Trends in Cognitive Sciences* 8(1), 8–11.
Garside, Roger. 1987. The CLAWS word-tagging system. In Roger Garside, Geoffrey Leech & Geoffrey Sampson (eds.), *The computational analysis of English: A corpus-based approach*, 30–41. London: Longman.
Greenbaum, Sidney & Jan Svartvik. 1990. The London–Lund Corpus of Spoken English. In Jan Svartvik (ed.), *The London–Lund Corpus of Spoken English: Description and research*, 11–59. Lund: Lund University Press.
Gries, Stefan Th. & Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide & James Pustejovsky (eds.), *Handbook of linguistic annotation*, 379–409. Dordrecht: Springer.
Halliday, M. A. K. 1989. *Spoken and written language*. Oxford: Oxford University Press.
Hardie, Andrew. 2014. Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal* 38(1), 73–103.
Hirst, Daniel. 2013. Anonymising long sounds for prosodic research. In Brigitte Bigi & Daniel Hirst (eds.), *Tools and resources for the analysis of speech prosody*, 36–7. Aix-en-Provence: Laboratoire Parole et Langage.
ICE: *International Corpus of English*. http://ice-corpora.net/ice/index.html (accessed 9 April 2020).
InqScribe. 2005–20. www.inqscribe.com (accessed 9 April 2021).
Johansson, Stig, Geoffrey Leech & Helen Goodluck. 1978. *Manual of information to accompany the Lancaster–Oslo/Bergen Corpus of British English, for use with digital computers*. Oslo: Department of English, University of Oslo. http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM (accessed 1 November 2020).
Leech, Geoffrey. 2007. New resources, or just better old ones? The Holy Grail of representativeness. In Marianne Hundt, Nadja Nesselhauf & Caroline Biewer (eds.), *Corpus linguistics and the web*, 133–49. Amsterdam: Rodopi.
Linell, Per. 2009. *Rethinking language, mind, and world dialogically: Interactional and contextual theories of human sense-making*. Charlotte, NC: Information Age Publishing.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina & Tony McEnery. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22(3), 319–44.
Meurer, Paul. 2012. Corpuscle – a new corpus management platform for annotated corpora. In Gisle Andersen (ed.), *Exploring newspaper language: Using the web to create and investigate a large corpus of modern Norwegian*, 29–50. Amsterdam: John Benjamins.
Nelson, Gerald. 2002. *Markup manual for spoken texts*. www.ice-corpora.uzh.ch/dam/jcr:72c70d5a-8da8-496f-b8dc-5fb66986c87c/spoken.pdf (accessed 25 February 2021).
Nelson, Gerald, Sean Wallis & Bas Aarts. 2002. *Exploring natural language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.
Ochs, Elinor. 1979. Transcription as theory. In Elinor Ochs & Bambi B. Schieffelin (eds.), *Developmental pragmatics*, 43–72. New York: Academic Press.
Põldvere, Nele. 2019. What's in a dialogue? On the dynamics of meaning-making in English conversation. PhD dissertation, Lund University.
Põldvere, Nele, Johan Frid, Victoria Johansson & Carita Paradis. 2021. Challenges of releasing audio material for spoken data: The case of the London–Lund Corpus 2. *Research in Corpus Linguistics* 9(1), 35–62.
Põldvere, Nele, Victoria Johansson & Carita Paradis. 2021. Resonance in dialogue: The interplay between intersubjective motivations and cognitive facilitation. Unpublished MS.
Põldvere, Nele, Victoria Johansson & Carita Paradis. In press. *A guide to the London–Lund Corpus 2 of spoken British English*. Lund Studies in English. Lund: Centre for Languages and Literature, Lund University.
Põldvere, Nele & Carita Paradis. 2019. Motivations and mechanisms for the development of the reactive *what-x* construction in spoken dialogue. *Journal of Pragmatics* 143, 65–84.
Põldvere, Nele & Carita Paradis. 2020. 'What and then a little robot brings it to you?' The reactive *what-x* construction in spoken dialogue. *English Language and Linguistics* 24(2), 307–32.
Schiffrin, Deborah. 1994. *Approaches to discourse*. Oxford: Blackwell.
Schmidt, Thomas. 2016. Good practices in the compilation of FOLK, the Research and Teaching Corpus of Spoken German. *International Journal of Corpus Linguistics* 21(3), 396–418.
Scott, Mike. 2020. *Wordsmith Tools*, version 8. lexically.net/wordsmith/ (accessed 1 November 2020).
Sinclair, John. 2005. Corpus and text: Basic principles. In Wynne (ed.), n.p.
Svartvik, Jan, Mats Eeg-Olofsson, Oscar Forsheden, Bengt Oreström & Cecilia Thavenius. 1982. *Survey of Spoken English: Report on research 1975–81*. Lund Studies in English 63. Lund: CWK Gleerup.
Svartvik, Jan & Randolph Quirk (eds.). 1980. *A corpus of English conversation*. Lund Studies in English 56. Lund: CWK Gleerup.
Sweney, Mark. 2019. Britons hang up the landline as call volumes halve. www.theguardian.com/business/2019/jan/05/britons-hang-up-landline-call-volumes-halve (accessed 1 November 2020).
Thompson, Paul. 2005. Spoken language corpora. In Wynne (ed.), n.p.
Wallis, Sean. 2006. *The International Corpus of English Corpus Utility Program* (ICECUP), version 3.1. London: Survey of English Usage, UCL. www.ucl.ac.uk/english-usage/resources/icecup/ (accessed 1 November 2020).
Weisser, Martin. 2017. Annotating the ICE corpora pragmatically: Preliminary issues & steps. *ICAME Journal* 41(1), 181–214.
Wynne, Martin. 2005a. Archiving, distribution and preservation. In Wynne (ed.), n.p.
Wynne, Martin (ed.). 2005b. *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books. http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf (accessed 1 November 2020).