RiCL Research in Corpus Linguistics

Challenges of releasing audio material for spoken data: The case of the *London-Lund Corpus 2*

Nele Põldvere[a/b] – Johan Frid[a] – Victoria Johansson[a] – Carita Paradis[a]
Lund University[a] / Sweden
University of Oslo[b] / Norway

Abstract – This article aims to describe key challenges of preparing and releasing audio material for spoken data and to propose solutions to these challenges. We draw on our experience of compiling the new *London-Lund Corpus 2* (LLC-2), where transcripts are released together with the audio files. However, making the audio material publicly available required careful consideration of how to, most effectively, 1) align the transcripts with the audio and 2) anonymise personal information in the recordings. First, audio-to-text alignment was solved through the insertion of timestamps in front of speaker turns in the transcription stage, which, as we show in the article, may later be used as a valuable complement to more robust automatic segmentation. Second, anonymisation was done by means of a *Praat* script, which replaced all personal information with a sound that made the lexical information incomprehensible but retained the prosodic characteristics. The public release of the LLC-2 audio material is a valuable feature of the corpus that allows users to extend the corpus data relative to their own research interests and, thus, broaden the scope of corpus linguistics. To illustrate this, we present three studies that have successfully used the LLC-2 audio material.

Keywords – audio-to-text alignment; anonymisation; corpus compilation; spoken corpora; prosody; *Praat*

1. INTRODUCTION[1]

With the advent of several new spoken corpora, challenges related to the various aspects of spoken corpus compilation are currently receiving more and more attention in the research community (e.g., Andersen 2016; Diemer *et al*. 2016; Kirk 2016; Sauer and

[1] We would like to express our gratitude to Bas Aarts and Sean Wallis from the *Survey of English Usage* (University College London) for giving us access to the LLC-1 audio material, and to the two anonymous reviewers, the editors of this special issue, and the general editors of *RICL* for their insightful comments on an earlier version of the manuscript. We are also grateful to *Lund University Humanities Lab*. This work has in part been funded by an infrastructure grant from the Swedish Research Council (Swe-Clarin, 2019–2024; contract no. 2017-00626). The compilation of LLC-2 has largely been 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.

Lüdeling 2016; Weisser 2017). However, these studies tend to focus on the part of corpora that constitutes the machine-readable data for spoken corpus research, that is, the transcriptions, rather than on the primary data from which the transcriptions have been derived, that is, the original audio recordings. The aim of this article is to describe and propose solutions to key challenges of preparing and releasing audio material for spoken data. It is based on our experience of compiling the new *London-Lund Corpus 2* (LLC-2; Põldvere *et al*. in press b.; see also the user guide in Põldvere *et al*. in press a.). LLC-2 is a half-a-million-word corpus of spoken British English dating from 2014 to 2019, and its compilation followed the same design criteria as in the world's first spoken corpus, the *London-Lund Corpus of Spoken English* (LLC-1) with data from the 1950s to the 1980s (see Section 3.1). In contrast to many other widely used spoken corpora in English, the transcripts in LLC-2 are released together with the audio files. However, for this to be possible, we had to tackle two major challenges: 1) the alignment of the transcripts with the audio files and 2) the anonymisation of personal information in the recordings. First, audio-to-text alignment was necessary in order to allow users to easily find relevant sections of the transcripts in the audio files and to improve the usability of LLC-2. The choice was between sophisticated automatic segmentation techniques and the simpler alternative of inserting timestamps during transcription. In this article, we explain why we decided to opt for the latter option and demonstrate the feasibility of combining it with more robust automatic segmentation (see Section 3.2). Second, the anonymisation of the audio recordings was mandatory out of respect for the speakers' privacy and legal protection of personal data. This procedure was, however, not straightforward because it required careful manipulation of the speech signal. We describe and explain why and how we anonymised the LLC-2 audio recordings using a *Praat* script developed by Hirst (2013) (see Section 3.3).

The benefits of releasing the LLC-2 audio material to the research community are immense. As is the case in many other spoken corpora, the transcriptions in LLC-2 are orthographic and contain information about basic features of spoken interaction such as pauses, overlapping speech and nonverbal vocalisations, but not prosodic and temporal information about pitch movement and the length of transitions between speaker turns. These features are, however, important for spoken language research because they carry useful information about speaker intent. Moreover, having access to prosodic and temporal information about speech broadens the field of corpus linguistics to go beyond

the traditional areas of lexicology, morphology, syntax and discourse analysis. With the release of the LLC-2 audio material, users can pursue these interests and extend the transcriptions using different speech analysis and annotation tools. To illustrate this, we provide examples of previous research on data from LLC-2 where the audio material was successfully used to carry out prosodic and temporal investigations of spoken interaction (see Section 3.4). Section 2 provides the background information.

2. AUDIO MATERIAL IN SPOKEN CORPORA

In this section, we will first present the core practices of how speech is represented in spoken corpora, and how these practices have influenced research conducted in two areas of linguistic inquiry: prosody and turn-taking (Section 2.1). Then, we review five well-known corpora of spoken British English and the extent to which they have made available the original audio material to facilitate more thorough investigations of the prosodic and temporal aspects of spoken interaction (Section 2.2).

2.1. Representations of speech in corpus linguistics

Compiling a spoken corpus is a complex and time-consuming task that requires careful decision-making at each stage of the process. Perhaps the most well-documented stage is the transcription stage, where the speech is turned into written form to provide the machine-readable material for browsing, searching and counting chunks in the corpus (e.g., Ochs 1979; Du Bois 1991; Crowdy 1994; Edwards 1995; Andersen 2016). To add value, the transcriptions may be complemented with layers of markup and annotation that convey additional information about the original speech event (e.g., Edwards 1995; Leech 2004; Kirk 2016; Sauer and Lüdeling 2016; Gries and Berez 2017; Weisser 2017). While corpus markup contains information about structural features inherent in speech production —such as who speaks, when and for how long— the function of corpus annotation is to add to the transcriptions linguistic information about, for example, parts-of-speech and syntactic parsing (Kirk and Andersen 2016: 291–292). The level of detail of the transcription, markup and annotation schemes adopted in spoken corpus projects depends on, among many other factors, the intended future uses of the corpus. Most of these uses tend to fall into the traditional areas of corpus linguistics such as lexicology, morphology, syntax and discourse analysis.

A much less well-documented stage of spoken corpus compilation is the process of making available the primary data from which the transcriptions have been derived, namely the original audio recordings (see, however, Diemer *et al*. 2016; Sauer and Lüdeling 2016; Schmidt 2016; Hoffmann and Arndt-Lappe submitted). This stage is, however, important because even the most detailed transcription, markup and annotation schemes lose valuable information about the original speech event in the transfer of the data to written form. Thus, the release of the audio material alongside the transcripts has the potential of extending corpus linguistics in new directions, that is, where the exploration of additional spoken features can add to our understanding of how spoken interaction works. In this article, we focus on two areas where this may prove useful: prosody and turn-taking.

Prosody is an essential component of human communication. Every utterance in spoken interaction contains prosodic features that convey important information about speaker intent. For example, the same expression has different interpretations depending on whether it receives a falling or rising intonation (compare *r\ight* as an expression of agreement and *r/ight* as a confirmation-seeking question).[2] Prosody research draws on data either from controlled laboratory experiments or speech corpora designed specifically for prosodic analyses (e.g., the *IViE Corpus of English Intonation in the British Isles*; see Grabe 2004).[3] Accordingly, the availability and quality of audio files are of utmost importance as "the research for which they are used is frequently focused on the speech signal itself" (Wichmann 2008: 188). This is different from corpus linguistics where, normally, corpora are intended to be useful for a wide variety of linguistic interests, and where many researchers consider the primary data to be the transcriptions with annotations of lexical, morpho-syntactic and discourse features (Oostdijk and Boves 2008: 196).

Turn-taking is a basic mechanism of dialogic spoken interaction and one of the main foci of Conversation Analysis (CA). Similar to corpus linguists, conversation analysts base their analyses on recordings of naturally occurring speech; however, most conversation analysts collect and transcribe their own data (Hoey and Kendrick 2017: 155) in order to ensure that the transcriptions are detailed enough to permit meaningful analyses for their purposes. For example, CA transcripts contain detailed information

[2] In the first instance, \ indicates a falling intonation contour from a high accented syllable and, in the second instance, / indicates a rising intonation contour from a low accented syllable.
[3] http://www.phon.ox.ac.uk/files/apps/IViE

about the boundaries of overlapping speech and the length of gaps between speaker turns in milliseconds. This information is important for understanding speaker intent because turns produced after a noticeable gap (after, say, 600 ms) have been found to signal interactional trouble (Roberts *et al*. 2006) and may be interpreted as "the first move toward some form of disagreement/rejection" (Clayman 2002: 235). The level of detail needed to transcribe the recordings means that the datasets in CA are relatively small, which goes well with the qualitative focus of the framework. More recent quantitative work, however, has also consulted larger corpora. Roberts *et al*. (2015), for example, used the *NXT-format Switchboard Corpus* (Calhoun *et al*. 2010), which includes detailed temporal chunking of phonetic segments and words, to automatically estimate the duration of transitions between speaker turns. Yet other quantitative studies in CA have made use of various speech analysis and annotation tools to manually identify beginnings and ends of speaker turns (e.g., *Praat* in Kendrick and Torreira 2015). Thus, analyses of the organisation of turn-taking in spoken interaction rely heavily on the availability either of richly annotated transcripts or the original audio material or both. However, as we will show in Section 2.2, it is not common that these features are available in spoken corpora, let alone the possibility to combine the transcripts with the audio to facilitate even more thorough analyses of turn-taking and prosody in spoken interaction.

2.2. A review of corpora of spoken British English

In this section, we review five well-known corpora of spoken British English and the extent to which they give access to the original audio material. The corpora are: 1) the spoken component of the first *British National Corpus* (Spoken BNC1994; cf. BNC Consortium 2007),[4] 2) the spoken component of the second *British National Corpus* (Spoken BNC2014; cf. Love *et al.* 2017),[5] 3) the *British Component of the International Corpus of English* (ICE-GB; cf. Nelson *et al.* 2002),[6] 4) the first *London-Lund Corpus* (LLC-1; Greenbaum and Svartvik 1990)[7] and 5) the second *London-Lund Corpus* (LLC-2; cf. Põldvere *et al*. in press b.).[8] Spoken BNC1994 and Spoken BNC2014 are

[4] http://www.natcorp.ox.ac.uk
[5] http://corpora.lancs.ac.uk/bnc2014
[6] http://ice-corpora.net/ice/index.html
[7] http://icame.uib.no
[8] https://projekt.ht.lu.se/llc2

large, multi-million-word corpora recorded in the early 1990s and 2010s, respectively. The remaining corpora are considerably smaller with approximately half-a-million words each. ICE-GB contains data from the 1990s, while LLC-1 was recorded as early as in the 1950s–1980s and LLC-2 was recorded as recently as 2014–2019. The corpora were selected for the review because they all provide access to spontaneous everyday conversation (either as part of the corpus or in full), which is the most rewarding conversational setting for studies of prosody and turn-taking, and they are available either for free or after payment of a licence fee.

Table 1 below presents basic information about how the corpora were transcribed, marked up and annotated to facilitate prosodic and temporal analyses of spoken interaction, and the availability of audio material in the corpora. The idea is to determine whether users can carry out analyses of the topics if they only have access to the transcripts, and, if not, what options there are for them to consult the original audio recordings.

As can be seen in Table 1, the general approach to transcription in the corpora is to adopt an enhanced orthographic transcription scheme, which involves a transcription of words enhanced by markups and annotations of basic spoken features such as pauses, overlapping speech, nonverbal vocalisations (e.g., laughter), etc. However, most of the corpora (i.e. Spoken BNC1994 and, to a lesser extent, Spoken BNC2014) contain only limited prosodic annotation, such as rough indications of pitch contours, or none at all (ICE-GB[9] and LLC-2). The main reasons why orthographic transcriptions take precedence in spoken corpora are because they are easier and less costly to implement than prosodic transcriptions, and because orthographic transcriptions are sufficient for a wide variety of corpus linguistic studies (Atkins *et al*. 1992: 10; Love *et al*. 2017: 334).

[9] It should be noted that *Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland* (SPICE-Ireland; cf. https://johnmkirk.etinu.net/cgi-bin/generic?instanceID=11), the pragmatically annotated version of the Irish component of the *International Corpus of English*, has been annotated for pitch location and direction (Kirk 2016).

Corpus	Transcription, markup and annotation			Audio material
	General	Prosody	Turn-taking	
Spoken BNC1994 (10 million words).	Enhanced orthographic transcription.	Little prosodic annotation (e.g., question marks are used to indicate *questioning utterances*).	Distinction between short (<5s) and long gaps; boundaries, but not length, of overlaps are marked.	Downloadable WAV files available from Audio BNC for free; audio playback of query matches available from the free online interface BNCweb; not all recordings included; subset of the recordings published on *Corpuscle*[10] (cf. Meurer 2012) as part of *The Bergen Corpus of London Teenage Language* (COLT), cf. Stenström *et al.* (1998).[11]
Spoken BNC2014 (11 million words).	Enhanced orthographic transcription.	Only questions with obvious rising intonation are marked.	Distinction between short (<5s) and long gaps; only presence/absence of overlaps is marked.	No public access to audio material; plans to anonymise and release the recordings.
ICE-GB (600,000 words).	Enhanced orthographic transcription.	No prosodic annotation.	Distinction between short (one syllable) and long gaps; boundaries, but not length, of overlaps are marked.	Audio playback of the recordings available at a cost from the *UCL Survey of English Usage*.
LLC-1 (500,000 words).	Prosodic and paralinguistic transcription.	Extensive prosodic annotation (e.g., tone units, nuclear tones, stress).	Distinction between short (one syllable) and long gaps; boundaries, but not length, of overlaps are marked.	No public access to audio material.
LLC-2 (500,000 words).	Enhanced orthographic transcription.	No prosodic annotation.	Only one type of gap is included (one syllable or longer); boundaries, but not length, of overlaps are marked.	Downloadable WAV files available from the *Lund University Humanities Lab*'s corpus server; all recordings included.[12]

Table 1: The comparison of the nature of transcriptions and the availability of audio material of five well-known corpora of spoken British English

The only corpus in Table 1 that contains detailed prosodic and paralinguistic transcriptions is LLC-1. The corpus is annotated for prosodic features such as tone unit boundaries, the direction of the nuclear tone, varying degrees of stress, and paralinguistic features such as whisper and creak (Svartvik and Quirk 1980; Greenbaum and Svartvik 1990). The prosodic annotations have provided searchable data for a broad range of corpus linguistic studies (e.g., Stenström 1984; Aijmer 1996; Paradis 1997; Altenberg 1998; Lenk 1998; Kaufmann 2002; Romero-Trillo 2014; Põldvere *et al.* 2016; Kimps 2018; Lin 2018). However, with data from the 1950s to the 1980s, LLC-1

[10] https://clarino.uib.no/korpuskel/page
[11] http://korpus.uib.no/icame/colt/
[12] Only one 10-minute university lecture is unavailable as per a request from the lecturer.

is less suited for contemporary investigations of speech. This is because prosodic alterations and variants have been found to go hand in hand with meaning shifts and change (Paradis 2008; Wichmann *et al*. 2010; Wichmann 2011; Põldvere and Paradis 2019, 2020), and the prosodic patterns found in English some 50 years ago may not be the same as in contemporary speech. Furthermore, the annotations in LLC-1 are based on auditory analysis, which is heavily reliant on subjective impressions (cf. Wichmann 2008: 202). Therefore, users may want to inspect the original speech signal to reinforce or counter auditory impressions and, thus, obtain more reliable results (see Section 3.4).

Investigations of turn-taking in the corpora in Table 1 are facilitated to the extent that all of them are annotated for whether the transition between the speaker turns is a gap or an overlap.[13] Many of the corpora have made available additional information such as distinctions between short and long gaps, and the boundaries of the overlapping speech, but none of them has gone as far as to measure the length of time between the speaker turns, as is commonly the case in CA (see Section 2.1 above). Thus, Table 1 shows that, while all the corpora facilitate rough analyses of the organisation of turn-taking, they are less well-suited for thorough investigations of the timing of turns in conversation.

When we compare the transcription schemes to the availability of the audio material, it becomes clear that the shortcomings of the transcriptions are not always compensated for by access to the original audio recordings or the access is in some way restricted. This explains, at least partly, why prosody and turn-taking —both of which are heavily dependent on the availability of the original speech signal— are conspicuously under-researched in corpus linguistics. The corpora that do not provide any kind of public access to the audio material are Spoken BNC2014 and LLC-1.[14] The main reason for this is that the recordings have not been anonymised and therefore cannot be publicly released (e.g., Love *et al*. 2017: 335; see also Section 3.3). The ICE-GB audio material is available via audio playback from the *Survey of English Usage* at University College London, which means that users can search for an expression in the

[13] Note that, for current purposes, we use the term ‘gap’ to refer to what are more commonly known in corpus annotation as ‘pauses’; however, they are a special kind of pauses in that they only occur between speaker turns.

[14] According to *UCL Survey of English Usage* (2020), the *Diachronic Corpus of Present-Day Spoken English* (cf. https://www.ucl.ac.uk/english-usage/projects/dcpse/), of which LLC-1 is part, only contains the orthographic transcriptions and not the original audio files. In the early days, researchers had to travel to the *Survey of English Usage* in London to be able to listen to the recordings. Many researchers today have access to the digital files; however, no systematic access has been provided to date.

corpus and listen to the passage containing that expression (*UCL Survey of English Usage* 2020; see also Wallis *et al*. 2006). However, this feature of ICE-GB is only available after payment of a licence fee, which together with the transcripts and the software for searching the corpus may amount to as much as £600–800 for an individual, single-copy licence.[15] Access to the Spoken BNC1994 audio material is free of charge. Moreover, users can choose between two formats: 1) the complete WAV audio files are available for download from Audio BNC (Coleman *et al*. 2012), and 2) the BNCweb online interface allows users to play back, as well as download, the audio of the query match and its immediate context (Hoffmann *et al*. 2008; Hoffmann and Arndt-Lappe submitted). The only downside is that neither Audio BNC nor BNCweb provides access to the complete dataset. According to Coleman *et al*. (2012: para. 2), "[t]here is a substantial number of XML transcription files for which we may no longer have the original audiotapes […] we also have quite a few recordings that we haven't yet related to any transcription." Moreover, for copyright reasons, neither of the audio editions of Spoken BNC1994 gives access to the recordings of a subset of BNC1994, namely COLT, which instead are published on the online interface *Corpuscle* via audio playback (for more information on *Corpuscle*, see Section 4). Coleman *et al*. (2012) estimate the size of the missing dataset in Audio BNC (and, by extension, BNCweb) to be around 2.5 million words. As we will show in Section 3.4, this is enough to pose problems for those who wish to use the audio material in their research.

In our work with the design and compilation of LLC-2, we decided to address the above-mentioned shortcomings and provide access to the complete set of recordings, which are time-aligned with the transcripts and anonymised to adhere to ethical standards (see Section 3 for details). The recordings can be accessed from the *Lund University Humanities Lab*'s corpus server as downloadable WAV files. We decided to make the LLC-2 audio material publicly available to allow users to extend the orthographic transcriptions relative to their own research interests using any of the free software available for annotating and analysing spoken data. However, preparing the audio files for release did not come without its challenges, which are the same challenges that have discouraged or prevented many corpus developers before us from doing it. The next section focuses on how we tackled these challenges and, thus,

[15] The prices are as of April 2021.

facilitated the investigation of prosodic and temporal aspects of spoken interaction in LLC-2 in subsequent research.

3. Challenges of preparing LLC-2 audio files for release

This section presents key challenges of making the LLC-2 audio material available to the research community. After a brief description of LLC-2 in Section 3.1, we examine the steps that we took to overcome two challenges of preparing the LLC-2 audio material for public release, audio-to-text alignment (Section 3.2) and anonymisation (Section 3.3). Section 3.4 presents three studies based on data from LLC-2 that demonstrate the usefulness of making the audio material publicly available.

3.1. LLC-2

As already mentioned in Section 1, LLC-2 is a half-a-million-word corpus of spoken British English dating from 2014 to 2019 (Põldvere *et al.* in press b.; see also the user guide in Põldvere *et al.* in press a.). It covers a range of discourse contexts including private contexts such as face-to-face conversation and phone/CMC conversation,[16] as well as public contexts such as broadcast media, parliamentary proceedings, spontaneous commentary, legal proceedings and prepared speech. In addition, efforts have been made to control for certain demographic categories such as the age and gender of the speakers. The size and design of LLC-2 are comparable to those of LLC-1 with data from the 1950s to the 1980s. As a result, LLC-2 can be used to study naturally occurring contemporary speech, on the one hand, and, on the other hand, it gives researchers the opportunity to make principled diachronic comparisons with LLC-1 of speech over the past half a century (see Section 3.4). The corpus will be released to the research community for free via the *Lund University Humanities Lab*'s corpus server in autumn 2021 (see also Section 4).[17] The release contains, among many other things, 184 XML-formatted transcription files and 183 audio files in WAV format.[18] In order to

[16] CMC = *Computer-Mediated Communication*.
[17] The corpus server can be accessed at https://www.humlab.lu.se/facilities/corpus-server
[18] In general, LLC-2 contains 100 texts, each around 5,000 words in size, with corresponding audio recordings, but since one text in the corpus can contain material from one recording only, or it can consist of multiple shorter recordings revolving around a similar subject matter and/or involving the same speaker(s), the total number of transcription and audio files is considerably higher.

facilitate the release of the audio material, we had to tackle two key challenges, which are discussed in the next two sections.

3.2. Audio-to-text alignment

The first key challenge was the alignment of the transcripts with the recordings. Audio-to-text alignment of this kind involves linking particular sections in the transcripts to the corresponding locations in the recordings in order to enhance the usability of the corpus. There are two broad options for how to deal with this (Thompson 2004). On the one hand, corpus developers may use highly sophisticated procedures for automatic alignment, which yield a best-fitting phonetic transcription of the audio and provide detailed timing information about all the vowels, consonants and words in the recordings. Such an approach was adopted in Spoken BNC1994, both in Audio BNC and BNCweb (Coleman *et al*. 2012; Hoffmann and Arndt-Lappe submitted). On the other hand, a simpler solution is to manually place markers in the transcripts to point to precise timings in the audio files. This functionality is often built into transcription software (e.g., ELAN; see Wittenburg *et al.* 2006) and it gets integrated into the transcription stage. In LLC-2, we adopted the latter approach. The reason for this was that the insertion of timestamps is easy to implement and provides sufficiently accurate points of entry into the audio files for a wide variety of corpus linguistic studies.

The tool used to insert timestamps in LLC-2 was *InqScribe* (2005–2020). *InqScribe* is a low-cost transcription software tool that enables users to perform all their transcriptions and audio playback in the same window. An important feature of the software is that it includes a simple functionality for inserting timestamps by means of customised keyboard shortcuts. In LLC-2, the insertion of timestamps was administered on a turn-by-turn basis. This means that, at the onset of each speaker turn in the recordings, a customised keyboard shortcut was used to launch a snippet containing the timestamp and the speaker's unique identifier. In recordings with only one speaker (e.g., prepared speech) or recordings with overly long contributions by one speaker (e.g., spontaneous commentary), timestamps were inserted every minute. The combination of the timestamps with the speakers' unique identifiers, inserted with one keyboard shortcut, meant that no extra time had to be spent on inserting the timestamps separately. Thus, this technique can be scaled up to larger corpora containing spontaneous everyday conversation, which, due to its messiness, still requires manual transcription (see McEnery 2018).

In order to facilitate compatibility with existing corpus tools, the *InqScribe* files were converted into canonical XML files. 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. Following the recommendations in Hardie (2014), we made additions to the standard set of XML tags where required. This is illustrated in the XML transcript in Figure 1 below, where each speaker turn is enclosed within the <turn> tag, which attributes for the number of the turn (*n*), the timestamp with the value format hh:mm:ss.ms, and, finally, the unique speaker identifier (*who*). The timestamps in LLC-2 help users find the appropriate places in the recordings with minimal effort, thus serving as valuable points-of-entry for more thorough analyses of the speaker turns. An obvious shortcoming of the XML transcripts is that they do not allow for immediate audio playback of the turns; however, we will facilitate this through the release of LLC-2 from an online interface (see Section 4 for details).

The availability of both the orthographic transcriptions and the corresponding audio recordings in LLC-2 also allows for the implementation of more sophisticated automatic alignment techniques to extend the use of the corpus to more areas. For example, for phonetic research it is usually desirable to have phonetic transcriptions as well as phonetically time-aligned boundaries between segments (Yuan *et al*. 2018). With a project of this scale, manual segmentation is not feasible as it is very costly in people-hours. Instead, automatic segmentation may be obtained through forced alignment. Forced alignment is the process of automatic alignment of an audio recording to a given transcript. Currently, the best systems for forced alignment make use of language-dependent dictionaries and acoustic models (Hosom 2009). The dictionaries are used to look up canonical phonetic representations of the words in the transcript, and the pre-trained acoustic models contain statistical representations of the acoustic information of the phonemes in language. The acoustic models analyse the audio recording, and the result is matched with the phonetic representation obtained from the dictionary in order to produce time-aligned segmentation. Some researchers have reported that there is a small decrease in accuracy compared to manual alignment (e.g., Hosom 2000). However, it is also the case that manual alignment by humans introduces a degree of random variability, while automatic alignment is rigorously systematic (see, e.g., Cosi *et al*. 1991; Baghai-Ravay *et al*. 2009). Weighing this in, the time gained from using automatic alignment is worth it.

```
<turn n="1" timestamp="00:00:00.17" who="S004">it's like some fitness place</turn>
<turn n="2" timestamp="00:00:02.15" who="S005">oh</turn>
<turn n="3" timestamp="00:00:02.25" who="S004">and some woman was just handing them out <pause/> but it looks alright they <overlap pos="start" n="1"/>do like <trunc>b</trunc><overlap pos="end" n="1"/></turn>
<turn n="4" timestamp="00:00:06.00" who="S005"><overlap pos="start" n="1"/><trunc>i</trunc> is it<overlap pos="end" n="1"/> all weird juices</turn>
<turn n="5" timestamp="00:00:07.24" who="S004">no they do like banana smoothies and stuff</turn>
<turn n="6" timestamp="00:00:10.19" who="S005">so yeah just weird juices <pause/> <overlap pos="start" n="2"/><vocal desc="laughs"/><overlap pos="end" n="2"/> well this could be nice</turn>
<turn n="7" timestamp="00:00:13.05" who="S004"><overlap pos="start" n="2"/><vocal desc="laughs"/><overlap pos="end" n="2"/></turn>
<turn n="8" timestamp="00:00:15.02" who="S004">any free stuff is <pause/> like I was just passing the woman with the fliers <pause/> and everyone was passing her by <pause/> and then I saw the flier said free and I was
<turn n="9" timestamp="00:00:23.10" who="S005">give <pause/></turn>
<turn n="10" timestamp="00:00:24.11" who="S004">yes <vocal desc="laughs"/> I'm <overlap pos="start" n="3"/>interested now<overlap pos="end" n="3"/></turn>
<turn n="11" timestamp="00:00:26.14" who="S005"><overlap pos="start" n="3"/><vocal desc="laughs"/><overlap pos="end" n="3"/> <pause/> Russell Square is it <overlap pos="start" n="4"/><vocal desc="gasps"/><overlap pos="e
<turn n="12" timestamp="00:00:30.22" who="S004"><overlap pos="start" n="4"/>yeah<overlap pos="end" n="4"/></turn>
<turn n="13" timestamp="00:00:34.02" who="S004"><overlap pos="start" n="5"/>no <vocal desc="laughs"/><overlap pos="end" n="5"/> <pause/> so now I have two free drinks I have the one from Eat which I haven't used yet <pa
<turn n="14" timestamp="00:00:44.00" who="S005">and the Waitrose one <pause/></turn>
<turn n="15" timestamp="00:00:45.03" who="S004">I haven't <overlap pos="start" n="6"/>got my Waitrose card yet<overlap pos="end" n="6"/></turn>
<turn n="16" timestamp="00:00:45.13" who="S005"><overlap pos="start" n="6"/>oh if if if you buy one<overlap pos="end" n="6"/> <pause/></turn>
<turn n="17" timestamp="00:00:47.19" who="S004"><vocal desc="laughs"/></turn>
<turn n="18" timestamp="00:00:50.12" who="S005">I'll catch up <pause/></turn>
<turn n="19" timestamp="00:00:51.20" who="S004">and <pause/> this one <pause/></turn>
<turn n="20" timestamp="00:00:55.20" who="S005">what do you have on tomorrow <pause/> you said you were really busy <pause/></turn>
<turn n="21" timestamp="00:00:58.15" who="S004">yes <pause/> I have <pause/> I have Icelandic <pause/> and then Spanish <pause/> so I have grammar for four hours <pause/></turn>
<turn n="22" timestamp="00:01:11.00" who="S005"><overlap pos="start" n="7"/>ugh<overlap pos="end" n="7"/></turn>
<turn n="23" timestamp="00:01:12.02" who="S004"><overlap pos="start" n="7"/>from nine till one<overlap pos="end" n="7"/></turn>
<turn n="24" timestamp="00:01:15.18" who="S005">you're gonna have to have a big lunch <pause/> are you gonna have time for breakfast with her <pause/></turn>
<turn n="25" timestamp="00:01:19.19" who="S004">I don't know <pause/></turn>
<turn n="26" timestamp="00:01:20.25" who="S005">cause you need to eat before that</turn>
<turn n="27" timestamp="00:01:22.03" who="S004">I missed it last time because I <overlap pos="start" n="8"/>was<overlap pos="end" n="8"/> late</turn>
<turn n="28" timestamp="00:01:23.24" who="S005"><overlap pos="start" n="8"/>what<overlap pos="end" n="8"/></turn>
<turn n="29" timestamp="00:01:24.14" who="S005">you missed it this morning</turn>
<turn n="30" timestamp="00:01:25.21" who="S004">no <overlap pos="start" n="9"/>I missed<overlap pos="end" n="9"/> it last <overlap pos="start" n="10"/>Thursday<overlap pos="end" n="10"/></turn>
<turn n="31" timestamp="00:01:26.10" who="S005"><overlap pos="start" n="9"/>oh<overlap pos="end" n="9"/></turn>
<turn n="32" timestamp="00:01:27.03" who="S005"><overlap pos="start" n="10"/>oh last last<overlap pos="end" n="10"/> time you didn't okay <pause/></turn>
<turn n="33" timestamp="00:01:31.18" who="S004">but then I ended up meeting her anyway</turn>
<turn n="34" timestamp="00:01:33.13" who="S005">uhuh <pause/> you don't have any like <trunc>biscuit</trunc> do you have any biscuits left</turn>
<turn n="35" timestamp="00:01:39.00" who="S004">yeah I have three shortbread <vocal desc="laughs"/> biscuits <overlap pos="start" n="11"/>left <pause/> I ate<overlap pos="end" n="11"/> all the rest of them</turn>
<turn n="36" timestamp="00:01:40.26" who="S005"><overlap pos="start" n="11"/><vocal desc="laughs"/> well<overlap pos="end" n="11"/></turn>
<turn n="37" timestamp="00:01:43.10" who="S005">aw well there's your breakfast <pause/> have <pause/> shortbread with Nutella and peanut butter and stuff like that</turn>
<turn n="38" timestamp="00:01:48.28" who="S004">for lunch I made <pause/> boiled eggs <overlap pos="start" n="12"/>eventually<overlap pos="end" n="12"/></turn>
<turn n="39" timestamp="00:01:52.00" who="S005"><overlap pos="start" n="12"/>mhm<overlap pos="end" n="12"/> <pause/> how many <pause/></turn>
<turn n="40" timestamp="00:01:54.13" who="S004">two <pause/></turn>
<turn n="41" timestamp="00:01:55.08" who="S005">mm <pause/></turn>
<turn n="42" timestamp="00:01:56.03" who="S004">I had to go down to the <pause/> to the kitchen on the next floor down <pause/></turn>
<turn n="43" timestamp="00:02:00.27" who="S005">how was it <pause/> was it like amazing <pause/></turn>
<turn n="44" timestamp="00:02:03.15" who="S004">well <pause/> the stuff worked <overlap pos="start" n="13"/>so that<overlap pos="end" n="13"/> was a change <vocal desc="laughs"/></turn>
<turn n="45" timestamp="00:02:05.25" who="S005"><overlap pos="start" n="13"/>mm<overlap pos="end" n="13"/></turn>
```

Figure 1: The illustration of an XML-formatted file in LLC-2

To illustrate the feasibility of forced alignment in LLC-2, we used the WebMAUS system (Schiel 1999; Kisler *et al*. 2017) to produce an alignment of the first few lines of the transcript in Figure 1 above and its corresponding audio recording (see also Sauer and Lüdeling 2016). The transcript and the recording are of a private and spontaneous face-to-face conversation. The result of the WebMAUS system is a TextGrid file, which can be used in the phonetics software *Praat* (Boersma 2001). This is illustrated in Figure 2 where the segmentations have been performed both at the level of words (upper annotation tier) and sounds (lower annotation tiers).

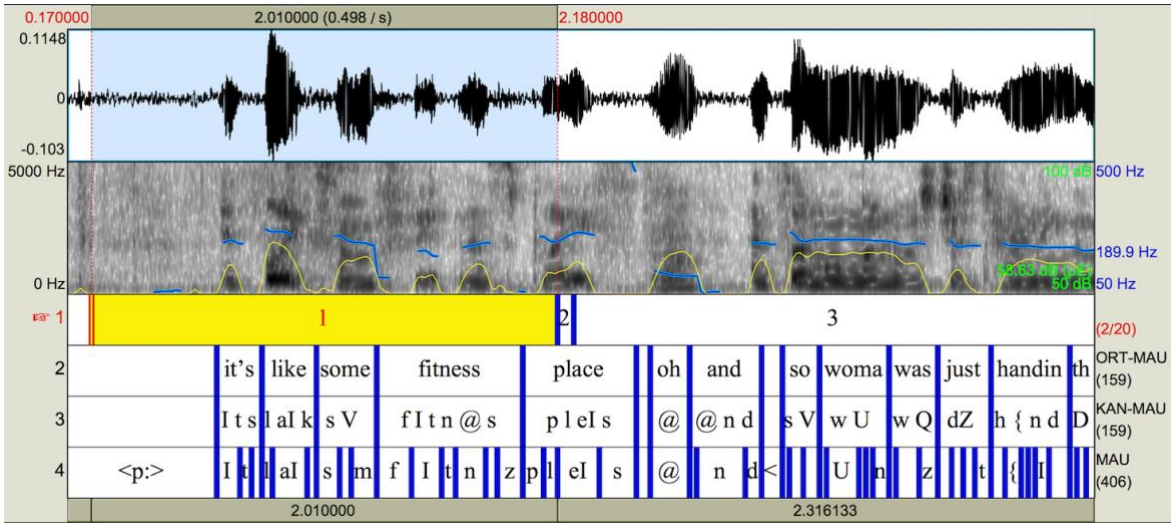

Figure 2: The output of the WebMAUS segmentation system in *Praat* based on the first few lines of the transcript in Figure 1

Looking at it qualitatively, the alignment of the speech signal and the phonetic segments in Figure 2 is very good. Admittedly, there are some misalignments for severely reduced and hasty speech, but that is to be expected in data of this kind. No quantitative evaluation has been made at this stage, as we have no ground truth data to evaluate the alignment against. One could use the automatic alignment as input for a manual correction procedure, which would be much faster than doing full transcription from scratch. Furthermore, the original timestamps in LLC-2 could be used to guide and improve manual editing of the segments. This may prove particularly useful in dealing with overlapping speech and background noise, which are notoriously difficult cases for forced alignment systems. Forced alignment is also highly sensitive to poor audio quality. LLC-2, too, contains private recordings that have been captured with speakers' personal smartphones (e.g., face-to-face conversation) or computer software (e.g., video conversation), which provide audio quality that is far from what phoneticians would consider ideal conditions for forced alignment. This said, we estimate that most of the

data in LLC-2 have been recorded with high-quality digital voice recorders, a feature that we expect to lead to a sufficiently high degree of segmentation accuracy. The alignment in Figure 2 (a private and spontaneous everyday conversation) is a case in point. Thus, looking forward, the prospect of generating for phonetic research automatic transcriptions in LLC-2 seems very promising.

3.3. Anonymisation

The second key challenge that we had to overcome when preparing the LLC-2 audio material for public release was the anonymisation of personal information in the recordings. Anonymisation is mandatory for any publicly available spoken corpus out of respect for the speakers' privacy in line with the *European Union's General Data Protection Regulation* (GDPR). It concerns the removal of all personal information that would allow an individual to be identified. In LLC-2, each speaker was assigned a unique identifier (e.g., <who="S004"> in Figure 1 above) and any references to people's names, addresses, phone numbers, etc., were removed, irrespective of whether these concerned the speakers themselves or any third parties not present in the conversation. The anonymisation was carried out on recordings obtained from private contexts, including 47 texts of face-to-face conversations, nine texts of phone/CMC conversations and two texts of university lectures, but no anonymisation was carried out on radio phone-ins or other types of recordings obtained from the public domain (e.g., podcast discussions).

The anonymisation of personal information during the transcription stage is relatively straightforward. In LLC-2, the transcribers were instructed to mark up all pieces of personal information by enclosing them within the <anon> tag, and to change the information while retaining the word class and number of syllables of the original (e.g., <anon>John</anon> for Sam). In this way, we were able to at least partly retain the socio-cultural information conveyed by the original proper name, including gender and, at times, also ethnicity (see Hasund 1998). A similar procedure was followed in the anonymisation of the transcriptions in ICE-GB and LLC-1 (e.g., Nelson 2002: 7).

The anonymisation of personal information in the original audio recordings is considerably more challenging. It requires careful manipulation of the speech signal, which, in turn, requires special training and adds considerably to the time and money

needed to release the corpus. For example, the reason why the Spoken BNC2014 audio material has not been publicly released yet is because the cost of anonymising the audio recordings went beyond the funding available for the project. However, additional funding will be sought to facilitate this in the future (Love *et al*. 2017: 335). Furthermore, the anonymisation techniques adopted in other spoken corpora have not been completely satisfactory, because they either make certain types of analyses impossible or they pose ethical problems. For example, the approach taken in Spoken BNC1994 consisted of locating and muting the portions of the audio recordings corresponding to the anonymisation tag. Such an approach, however, removes important prosodic information about the original speech signal. Other techniques retain the prosodic information but are problematic in ethical terms. Hirst (2013), for example, reviews two techniques commonly used in psycho-acoustic experiments: 1) the inversion of the spectrum of the speech signal, and 2) the application of a filter that removes the spectral information. However, the problem with those solutions is that, in the first instance, the second inversion of the spectrum restores the original speech signal, and, in the second instance, even quite severe filtering does not make the speech signal unintelligible.

The technique adopted in LLC-2 is based on a *Praat* script written and developed by Hirst (2013).[19] To the best of our knowledge, it has not been implemented in other similar corpora so far.[20] The script works on the basis that the portion of the speech signal that has been marked by the corpus developer with the keyword *buzz* is replaced by a *hum* sound that makes the lexical content of the signal incomprehensible but retains the pitch and intensity envelope of the original. The advantage of this technique is that it is reliable and retains linguistically useful information such as prosody. Moreover, running the script is relatively easy and can be achieved with only minimal training in *Praat*. A somewhat fortuitous side effect is that the task effectively produces data for building a named entity recognition system that can automatically find new portions (names, locations, etc.) that are possible candidates for being anonymised.

An illustration of how the *Praat* script works is given in Figures 3 and 4. Both figures represent the speech signal of a public recording in LLC-2, together with the

[19] The script is freely available at https://hdl.handle.net/11403/sldr000526/v6

[20] The script is currently used in *LangAge Corpora* (cf. http://www.uni-potsdam.de/langage/); however, the corpora are in French and contain specialised content of sociolinguistic interviews with elderly speakers only (Gerstenberg *et al*. 2017).

location and direction of the pitch contour (blue line) and the intensity profile (yellow line).[21] The audio snippet extracted from the recording contains the utterance *Jenni Rodd is a cognitive psychologist at University College London* in which the personal pieces of information are the name and workplace of the person talked about. In Figure 3, this information is marked with the keyword *buzz* to indicate the portions of the speech signal that will be anonymised. Figure 4 presents the end result where the information has been anonymised, and where the pitch and intensity envelopes are the same as in the original.

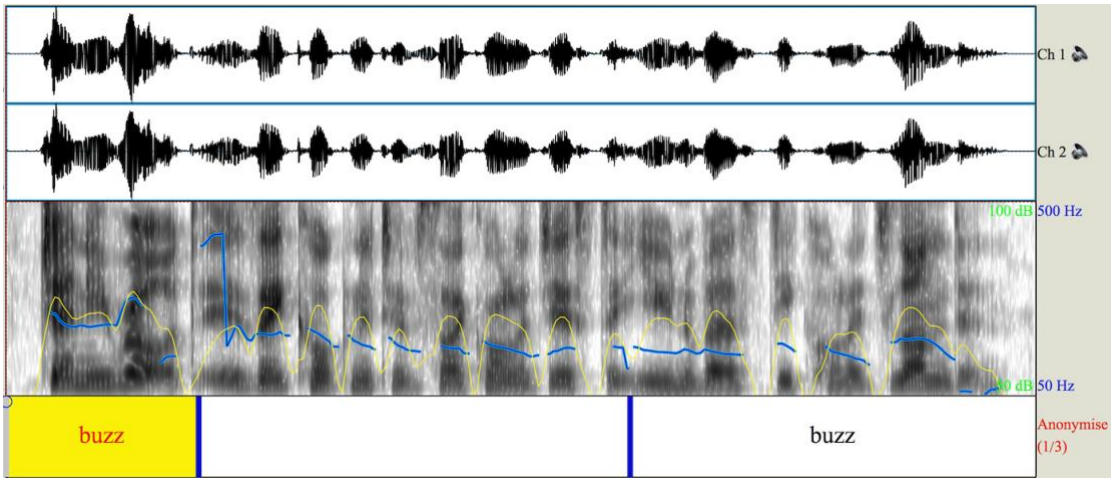

Figure 3: The original speech signal, pitch contour and intensity profile of the utterance *Jenni Rodd is a cognitive psychologist at University College London*. Click on the image to listen to the audio

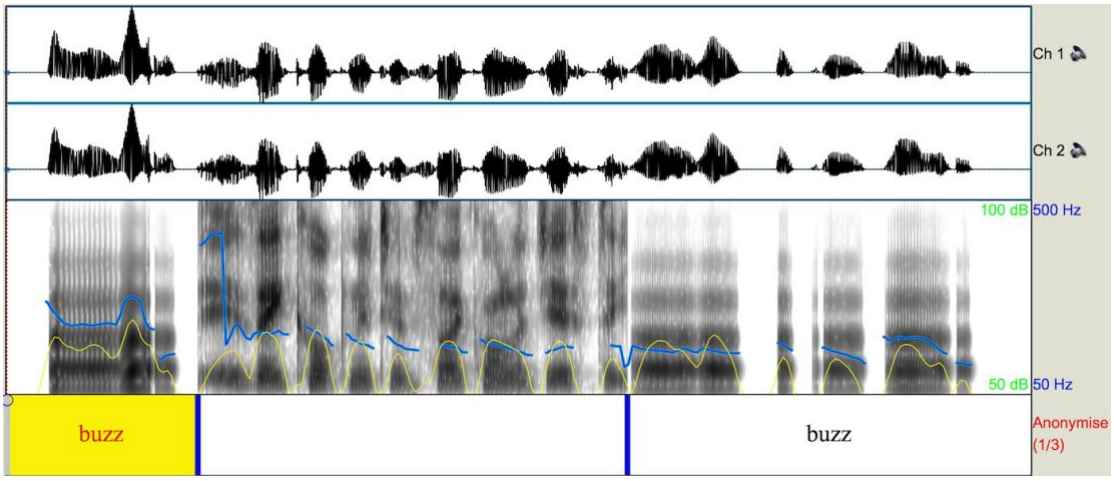

Figure 4: The manipulated speech signal, pitch contour and intensity profile of the utterance *Jenni Rodd is a cognitive psychologist at University College London*. Click on the image to listen to the audio[22]

[21] Note that since the recording, a podcast discussion, was obtained from the public domain, it has not been anonymised in the corpus.

[22] The audio snippets corresponding to the figures are also available at https://projekt.ht.lu.se/llc2/anonymisation.

In total, we anonymised approximately 1,300 personal pieces of information in LLC-2. The timestamps in the transcripts (see Section 3.2 above) helped us locate the information in the recordings with much less effort than if the transcripts had not been aligned with the recordings. This said, the manual nature of the task requires that corpus developers allow for a sufficient amount of time for completing it, which may prove impractical for larger corpora. However, the end result is worth the effort because it gives us a corpus that meets the ethical requirements of anonymity, which is mandatory for the public release of the audio material, and it also facilitates prosodic analyses on the corpus.

3.4. Applications of LLC-2 audio material

After tackling the challenges above, the LLC-2 audio material can be released to the public. The audio recordings are useful in a variety of areas in linguistics that, traditionally, have been outside the main focus of corpus linguistics. This section illustrates how the LLC-2 audio material can be used for investigations of the prosodic and temporal aspects of spoken interaction. It demonstrates three studies (Põldvere and Paradis 2019, 2020; Põldvere *et al*. submitted) based on data from LLC-2 that combined the orthographic transcriptions with instrumental analyses of the recordings to facilitate more thorough and, at times, even more reliable analyses of the phenomena in question.

Põldvere and Paradis (2019, 2020) were both concerned with a construction that previously had not received any attention in the literature, namely the reactive *what-x* construction. While Põldvere and Paradis (2020) set out to describe and define the constructional properties of the construction in LLC-2, Põldvere and Paradis (2019) tracked the development of the construction from LLC-1 to LLC-2, that is, over the past half a century. The LLC-1 audio material was made available to us by the *Survey of English Usage*. The analyses showed that the reactive *what-x* construction is a conventionalised construction in English that is characterised by a range of formal and functional properties that distinguish it from other, better-known *what*-constructions. One of these properties is prosody. Consider the utterance in bold in (1), which is an example of the reactive *what-x* construction in LLC-2.[23]

[23] Note that the transcriptions in this section have been slightly simplified in order to facilitate the task of the reader.

(1) <S051> I know it's ridiculous to plan Christmas already <pause/> although I did see <pause/> Christmas food in Sainsbury's yesterday

<S052> **what mince pies** <pause/>

<S051> all sorts of stuff

According to Põldvere and Paradis (2019, 2020), the reactive *what-x* construction always comprises the interrogative *what* and a subsequent complement, and its discursive meaning is to react to an immediately preceding turn to call it into question. In (1), *what* is followed by the noun phrase *mince pies*, used to react to the interlocutor's prior turn and to verify the specific Christmas food sold at Sainsbury's. However, an important property of the reactive *what-x* construction that cannot be derived from the orthographic transcription is that *what* always forms one and the same tone unit with the complement. This was determined in the studies through instrumental analyses of the construction in *Praat*.[24] Figure 5 illustrates the pitch contour of the reactive *what-x* construction in (1). As can be seen in the figure, *what* and *mince pies* form one and the same tone unit where *what* is realised as an unaccented pre-head of the unit, and the nuclear pitch accent, rise-fall, is on *pies*.[25] This information would have remained hidden to us had we not consulted the LLC-2 audio material.

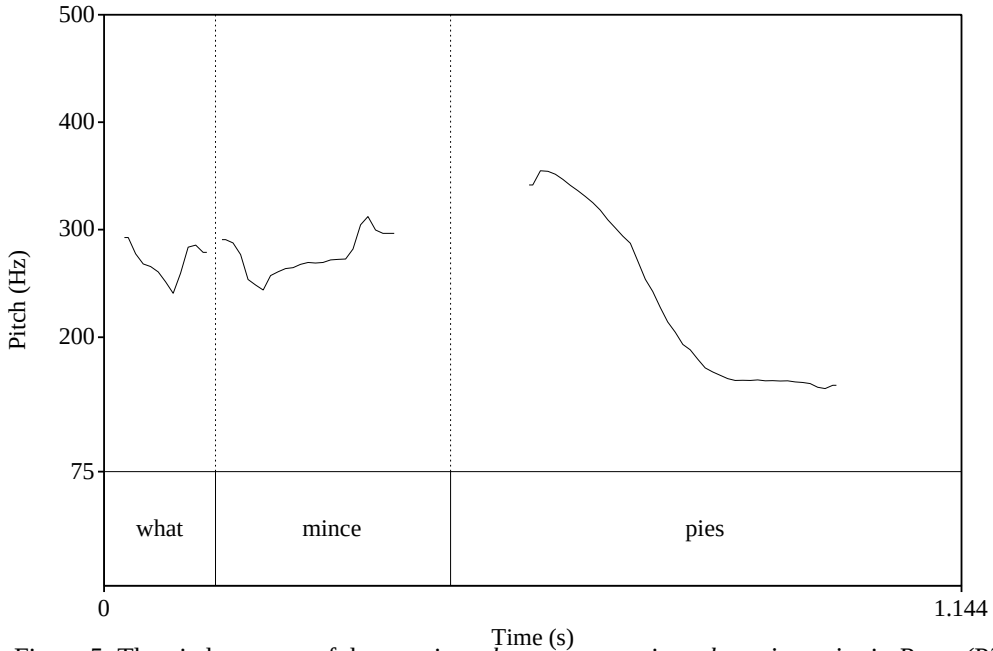

Figure 5: The pitch contour of the reactive *what-x* construction *what mince pies* in *Praat* (Põldvere and Paradis 2020: 320)

[24] In a few rare cases, the quality of the audio recordings was not good enough for instrumental analyses. In such cases, the recordings were auditorily inspected by both co-authors, and the decision as to the boundaries of the tone units and the types of nuclear pitch accents were made together.

[25] The prosodic analyses in Põldvere and Paradis (2019, 2020) follow the British tradition of intonation analysis where the basic unit is the tone, and where the direction of the pitch contour is a fall, rise, level, fall-rise or rise-fall (see, e.g., Cruttenden 1997).

Furthermore, the original audio recordings in the corpora helped us distinguish between the reactive *what-x* construction and a closely related *what*-construction, the pragmatic marker *what* (e.g., Brinton 2017). In many cases, the only property that sets the two constructions apart is that the pragmatic marker *what* always forms its own tone unit (e.g., *wh/\at # a b/\ird*),[26] which contributes to its interpretation as an expression of surprise and incredulity rather than a request for verification. Thus, the pragmatic marker *what* and the reactive *what-x* construction are two different constructions in English with distinct formal and functional characteristics. Without consulting the LLC-2 audio material, we would have missed this difference. In fact, this was a problem that we encountered in Põldvere and Paradis (2019), which included an additional analysis of the reactive *what-x* construction in Spoken BNC1994. Specifically, the missing audio data in the corpus meant that we were unable to classify eight per cent of the *what*-constructions included in the analysis. Furthermore, a comparison of the instrumental analysis of the LLC-1 audio material and the prosodic annotations revealed that not all instances of *what* in the transcripts had been assigned the correct prosodic pattern; in other words, what looked like the pragmatic marker *what* was in fact the reactive *what-x* construction, and vice versa. Thus, access to the LLC-1 audio material allowed us to validate the prosodic annotations against instrumental analyses and obtain more reliable results.

In Põldvere *et al*. (submitted), we used the LLC-2 audio material to investigate the timing of turns in conversational sequences where the speakers reproduce constructions from prior turns, called 'dialogic resonance' (Du Bois 2014). Consider the sequence in (2), taken from LLC-2, where the resonance is achieved through the speakers' choice of words and structures.

(2) <S002> yeah well so don't end up at home every day
<S003> I won't be at home every day <anon>Sara</anon>

According to Du Bois (2014), dialogic resonance emerges because speakers want to engage with the words of their interlocutors for various socio-communicative purposes. For example, previous work has showed that resonance is a fruitful way to express disagreement in spoken interaction (e.g., Dori-Hacohen 2017), as illustrated in (2). While Du Bois acknowledges the role of priming in resonance, this is not tested in his

[26] The hash sign (#) indicates a tone unit boundary between *what* and *a bird*, and /\ indicates a rising-falling pitch contour.

work. Instead, priming is the central mechanism of Garrod and Pickering's (2004) interactive alignment theory, which states that prior expression primes the reuse of the same linguistic representations by the next speaker. Thus, priming has a facilitating effect in resonance due to cognitive activation in the prior turn. In order to investigate the role of cognitive facilitation in resonance, we operationalised it as the time it takes for speakers to respond to the interlocutor's prior turn, based on the assumption that the timing of turns in conversation reflects the degree to which linguistic constructions are activated and accessible to the next speaker. The prediction was that transitions between speaker turns are faster in resonating sequences compared to when the turns are constructed from scratch. The results confirmed this prediction, showing that cognitive facilitation gives speakers the necessary tools to counter the temporal challenges of spontaneous conversation.

The analysis in Põldvere *et al*. (submitted) would not have been possible without the LLC-2 audio material. This is because the transcriptions in LLC-2 contain only limited information about turn transitions, showing whether a transition is a gap or an overlap but not its length in milliseconds. However, this information is crucial for systematic investigations of the timing of turns in conversation. In order to extract reliable measurements of turn transitions in the data, we used the multimodal annotation tool ELAN. The advantage of using ELAN over other speech analysis software such as *Praat* is that ELAN allows for the annotation of the speech signal using multiple tiers that can be created freely by the analyst. Moreover, the length of the annotations in milliseconds can be easily exported to a spreadsheet or database software for statistical analysis. Figure 6 illustrates the speech signal and the corresponding annotation of the conversational sequence in (2) above. As can be seen in the figure, the annotation scheme includes the orthographic transcription of the utterances in the conversational sequence, and the type of transition between the utterances, in this case a gap. The exported data reveal that the length of the gap is eight milliseconds, which is very fast considering that the dialogic function of the response is to express disagreement, a dispreferred response. The rest of the annotations in Figure 6 need not concern us here.

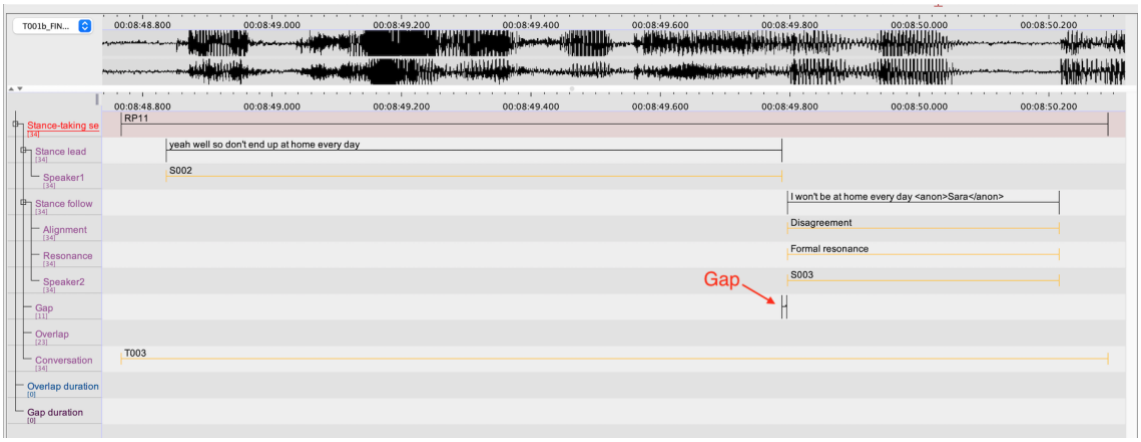

Figure 6: The illustration of a gap between the resonating utterances expressing disagreement, *yeah well so don't end up at home every day* and *I won't be at home every day <anon>Sara</anon>* in ELAN

In sum, the studies above show that, with access to the LLC-2 audio material, and the appropriate software, users have at their disposal all the necessary tools to carry out thorough and reliable analyses of prosody and turn-taking in spoken interaction, and therefore promote the extension of corpus linguistics in new directions. The cost and effort associated with overcoming the methodological challenges of preparing the audio material for public release has been a small price to pay for such a gain.

4. Conclusion and future work

The aim of this article has been to describe key challenges of preparing and releasing audio material for spoken data and to propose solutions to these challenges. We have focused on two challenges that we had to tackle during the compilation of LLC-2: 1) the alignment of the orthographic transcriptions with the audio files and 2) the anonymisation of personal information in the recordings. Audio-to-text alignment was necessary because it allows users to easily link relevant sections in the transcripts to the corresponding locations in the audio files. We opted for a solution that involved inserting timestamps by means of *InqScribe* in front of speaker turns to indicate to the users where each turn begins. As shown, this solution can be effectively combined with more sophisticated automatic segmentation techniques (e.g., the WebMAUS forced alignment system). The second challenge concerned the anonymisation of personal information in the audio recordings, which was mandatory in order to abide by the ethical and legal principles of privacy and data protection. For the best result possible, we used a *Praat* script developed by Hirst (2013). The script replaces all personal information in the recordings with a sound that makes the lexical information

incomprehensible but retains the prosodic characteristics of the original speech signal. The advantage of this technique over some of the other techniques suggested in the literature is that it is reliable and makes possible a wide variety of linguistic analyses, including prosody.

The release of the LLC-2 audio material together with the transcripts is unique because it opens up research opportunities that extend the scope of corpus linguistics in new and exciting directions. This article has focused on two areas that are conspicuously under-researched in spoken corpus research: prosody and turn-taking. Drawing on three studies based on data from LLC-2, we have demonstrated that the LLC-2 audio material can be used to perform thorough and reliable investigations of the prosodic and temporal aspects of spoken interaction using freely available speech analysis and annotation tools. In our view, the opportunities that the LLC-2 audio recordings offer for spoken corpus research overweigh the methodological challenges of making them publicly available. Therefore, future corpus developers are encouraged to factor in the time and effort of tackling these challenges. At the same time, we acknowledge that the techniques presented here may be more suitable for smaller-scale corpora such as LLC-2 rather than larger, multi-million-word national corpora. This is mainly due to the considerable amount of manual effort needed, particularly in the annotation of personal pieces of information in *Praat*. This said, the rapid technological advances in machine learning and audio-to-text technologies give us hope that, in the not-too-distant future, these techniques can be scaled up to larger corpora, too. In the meantime, the present techniques could be applied to a subset of a larger corpus in order to facilitate prosodic and temporal analyses on, at least, a part of it.

Future work on LLC-2 involves making the recordings and transcripts available from the free corpus management and analysis system *Corpuscle* (Meurer 2012). *Corpuscle* will enable the implementation of various corpus linguistic techniques on LLC-2, and the possibility to carry out restricted searches on the corpus data based on the many demographic categories available in the metadata. The release of LLC-2 from *Corpuscle* also means that users will no longer have to navigate the individual XML transcription files and WAV audio files to be able to listen to relevant sections of the transcripts. Instead, this process will be made considerably quicker by the audio playback function of *Corpuscle* in which case a click on the transcription immediately plays back the corresponding part of the recording. The most promising feature of

Corpuscle for LLC-2 is that the audio playback works on a turn-by-turn basis, meaning that the timestamps in the transcripts will be sufficient for setting it up. We hope that the combination of downloadable and time-aligned transcription and audio files with online audio snippets will lead to even more diverse uses of LLC-2 and facilitate seamless experiences of using the corpus.

REFERENCES

Aijmer, Karin. 1996. *Conversational Routines in English: Convention and Creativity*. London: Longman.

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word combinations. In Anthony P. Cowie ed. *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press, 101–122.

Andersen, Gisle. 2016. Semi-lexical features in corpus transcription: Consistency, comparability, standardisation. *International Journal of Corpus Linguistics* 21/3: 323–347.

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

Baghai-Ravay, Ladan, Greg Kochanski and John Coleman. 2009. Precision of phoneme boundaries derived using Hidden Markov Models. *Proceedings of INTERSPEECH 2009, Tenth Annual Conference of the Interantional Speech Communication Association*, 2879–2882.

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5/9–10: 341–345.

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. http://www.natcorp.ox.ac.uk. (9 April, 2021.)

Brinton, Laurel J. 2017. *The Evolution of Pragmatic Markers in English: Pathways of Change*. Cambridge: Cambridge University Press.

Calhoun, Sasha, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman and David Beaver. 2010. The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation* 44/4: 387–419.

Clayman, Steven E. 2002. Sequence and solidarity. In Shane R. Thye and Edward J. Lawler eds. *Group Cohesion, Trust and Solidarity*. Oxford: Elsevier, 229–253.

Coleman, John, Ladan Baghai-Ravary, John Pybus and Sergio Grau. 2012. Audio BNC: The Audio Edition of the Spoken British National Corpus. http://www.phon.ox.ac.uk/AudioBNC (24 February, 2020.)

Cosi, Piero, Daniele Falavigna and Maurizio Omologo. 1991. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. *Proceedings of EUROSPEECH 1991, Second European Conference on Speech Communication and Technology*, 693–696.

Crowdy, Steve. 1994. Spoken corpus transcription. *Literary and Linguistic Computing* 9/1: 25–28.

Cruttenden, Alan. 1997. *Intonation*. Cambridge: Cambridge University Press.

Diemer, Stefan, Marie-Louise Brunner and Selina Schmidt. 2016. Compiling computer-mediated spoken language corpora: Key issues and recommendations. *International Journal of Corpus Linguistics* 21/3: 348–371.

Dori-Hacohen, Gonen. 2017. Creative resonance and misalignment stance: Achieving distance in one Hebrew interaction. *Functions of Language* 24/1: 16–40.

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 and Jenny Thomas eds. *Spoken English on Computer: Transcription, Mark-Up and Application*. New York: Longman, 19–34.

Garrod, Simon and Martin J. Pickering. 2004. Why is conversation so easy? *TRENDS in Cognitive Sciences* 8/1: 8–11.

Gerstenberg, Annette, Valerie Hekkel, Julie Marie Kairet and Adélie Soumier-Vendé. 2017. *LangAge Corpora: Resources for Language and Aging Research*. Poster presentation at CLARe3, Berlin, Germany.

Grabe, Esther. 2004. Intonational variation in urban dialects of English spoken in the British Isles. In Peter Gilles and Jörg Peters eds. *Regional Variation in Intonation*. Tübingen: Niemeyer, 9–31.

Greenbaum, Sidney and Jan Svartvik. 1990. The *London-Lund Corpus of Spoken English*. In Jan Svartvik ed. *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press, 11–59.

Gries, Stefan Th. and Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide and James Pustejovsky eds. *Handbook of Linguistic Annotation*. Berlin: Springer, 379–409.

Hardie, Andrew. 2014. Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal* 38: 73–103.

Hasund, Ingrid Kristine. 1998. Protecting the innocent: The issue of informants' anonymity in the COLT corpus. In Antoinette Renouf ed. *Explorations in Corpus Linguistics*. Amsterdam: Rodopi, 13–28.

Hirst, Daniel. 2013. Anonymising long sounds for prosodic research. In Brigitte Bigi and Daniel Hirst eds. *Tools and Resources for the Analysis of Speech Prosody*. Aix-en-Provence: Laboratoire Parole et Langage, 36–37.

Hoey, Elliott M. and Kobin H. Kendrick. 2017. Conversation Analysis. In Annette M. B. de Groot and Peter Hagoort eds. *Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide*. Hoboken: Wiley-Blackwell, 151–173.

Hoffmann, Sebastian and Sabine Arndt-Lappe. Submitted. Better data for more researchers – Using the audio features of BNCweb.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee and Ylva Berglund Prytz. 2008. *Corpus Linguistics with BNCweb – A Practical Guide*. Frankfurt am Main: Peter Lang.

Hosom, John-Paul. 2000. *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. Hillsboro, OR: Oregon Health and Science University dissertation.

Hosom, John-Paul. 2009. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication* 51/4: 352–368.

InqScribe. 2005–2020. Computer software. https://www.inqscribe.com/ (9 April, 2021.)

Kaufmann, Anita. 2002. Negation and prosody in British English. *Journal of Pragmatics* 34/10: 1473–1494.

Kendrick, Kobin H. and Francisco Torreira. 2015. The timing and construction of preference: A quantitative study. *Discourse Processes* 52/4: 255–289.

Kimps, Ditte. 2018. *Tag Questions in Conversation: A Typology of their Interactional and Stance Meanings*. Amsterdam: John Benjamins.

Kirk, John M. 2016. The Pragmatic Annotation Scheme of the *SPICE-Ireland Corpus*. *International Journal of Corpus Linguistics* 21/3: 299–322.

Kirk, John M. and Gisle Andersen. 2016. Compilation, transcription, markup and annotation of spoken corpora. *International Journal of Corpus Linguistics* 21/3: 291–298.

Kisler, Thomas, Uwe Reichel and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45: 326–347.

Leech, Geoffrey. 2004. Adding linguistic annotation. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. http://users.ox.ac.uk/~martinw/dlc/chapter2.htm (24 February, 2020.)

Lenk, Uta. 1998. *Marking Discourse Coherence: Functions of Discourse Markers in Spoken English*. Tübingen: Gunter Narr Verlag.

Lin, Phoebe. 2018. *The Prosody of Formulaic Sequences: A Corpus and Discourse Approach*. London: Bloomsbury.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a corpus of everyday conversations. *International Journal of Corpus Linguistics* 22/3: 319–344.

McEnery, Tony. 2018. The Spoken BNC2014: The corpus linguistic perspective. In Vaclav Brezina, Robbie Love and Karin Aijmer eds. *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge, 10–15.

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*. Amsterdam: John Benjamins, 29–50.

Nelson, Gerald. 2002. *Markup Manual for Spoken Texts.* http://ice-corpora.net/ice/index.html (24 February, 2020.)

Nelson, Gerald, Sean Wallis and 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 and Bambi B. Schiefflen eds. *Developmental Pragmatics*. New York: Academic Press, 43–72.

Oostdijk, Nelleke and Lou Boves. 2008. Preprocessing speech corpora: Transcription and phonological annotation. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook* Vol. 1. Berlin: Mouton de Gruyter, 642–663.

Paradis, Carita. 1997. *Degree Modifiers of Adjectives in Spoken British English*. Lund: Lund University Press.

Paradis, Carita. 2008. Configurations, construals and change: Expressions of degree. *English Language and Linguistics* 12/2: 317–343.

Põldvere, Nele and 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 and 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: 307–332.

Põldvere, Nele, Matteo Fuoli and Carita Paradis. 2016. A study of dialogic expansion and contraction in spoken discourse using corpus and experimental techniques. *Corpora* 11/2: 191–225.

Põldvere, Nele, Victoria Johansson and Carita Paradis. In press a. *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, Victoria Johansson and Carita Paradis. In press b. On the *London-Lund Corpus 2*: Design, challenges and innovations. *English Language and Linguistics* 25/3.

Põldvere, Nele, Victoria Johansson and Carita Paradis. Submitted. Resonance in dialogue: The interplay between intersubjective motivations and cognitive facilitation.

Roberts, Felicia, Alexander L. Francis and Melanie Morgan. 2006. The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Communication* 48/9: 1079–1093.

Roberts, Seán G., Francisco Torreira and Stephen C. Levinson. 2015. The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology* 6: 1–16.

Romero-Trillo, Jesús. 2014. 'Pragmatic punting' and prosody. In María de los Ángeles Gómez González, Francisco José Ruiz de Mendoza Ibáñez, Francisco Gonzálvez-García and Angela Downing eds. *The Functional Perspective on Language and Discourse: Applications and Implications*. Amsterdam: John Benjamins, 209–222.

Sauer, Simon and Anke Lüdeling. 2016. Flexible multi-layer spoken dialogue corpora. *International Journal of Corpus Linguistics* 21/3: 419–438.

Schiel, Florian. 1999. Automatic phonetic transcription of non prompted speech. In John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville and Ashlee C. Baile eds. *Proceedings of ICPhS 1999, Fourteenth International Congress of Phonetic Sciences*, 607–610.

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.

Stenström, Anna-Brita. 1984. *Questions and Responses in English Conversation*. Malmö: Gleerup.

Stenström, Anna-Brita, Gisele Andersen, Kristine Hasund, Kristina Monstad and Hanne Aas. 1998. *User's Manual to Accompany The Bergen Corpus of London Teenage Language* (COLT). Bergen: Department of English, University of Bergen.

Svartvik, Jan and Randolph Quirk eds. 1980. *A Corpus of English Conversation*. Lund: Gleerup.

Thompson, Paul. 2004. Spoken language corpora. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. http://users.ox.ac.uk/~martinw/dlc/chapter5.htm (9 April, 2021.)

UCL Survey of English Usage. 2020. https://www.ucl.ac.uk/english-usage/ (5 April, 2021.)

Wallis, Sean, Gerald Nelson and Bas Aarts eds. 2006. *The British Component of the International Corpus of English* (ICE-GB), Release 2. London: *Survey of English Usage* computer software.

Weisser, Martin. 2017. Annotating the ICE corpora pragmatically – Preliminary issues & steps. *ICAME Journal* 41/1: 181–214.

Wichmann, Anne. 2008. Speech corpora and spoken corpora. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook* Vol. 1. Berlin: Mouton de Gruyter, 187–206.

Wichmann, Anne. 2011. Grammaticalization and prosody. In Bernd Heine and Heiko Narrog eds. *The Oxford Handbook of Grammaticalization*. Oxford: Oxford University Press, 331–341.

Wichmann, Anne, Anne-Marie Simon-Vandenbergen and Karin Aijmer. 2010. How prosody reflects semantic change: A synchronic case study of *of course*. In Kristin Davidse, Lieven Vandelanotte and Hubert Cuyckens eds. *Subjectification, Intersubjectification and Grammaticalization*. Berlin: Mouton de Gruyter, 103–154.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bente Maegaard, Joseph Mariani, Jan Odijk and Daniel Tapias eds. *Proceedings of LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 1556–1559.

Yuan, Jiahong, Wei Lai, Chris Cieri and Mark Liberman. 2018. Using forced alignment for phonetics research. In Chu-Ren Huang, Peng Jin and Shu-Kai Hsieh eds. *Chinese Language Resources and Processing: Text, Speech and Language Technology*. Springer.

Corresponding author
Nele Põldvere
Centre for Languages and Literature
Lund University
Box 201
221 00 Lund
Sweden
Email: nele.poldvere@englund.lu.se

received: January 2020
accepted: March 2021