The new London–Lund Corpus 2
Key methodological challenges and innovations
Pöldvere, Nele; Johansson, Victoria; Paradis, Carita

2020

Document Version:
Publisher's PDF, also known as Version of record

Link to publication

Citation for published version (APA):

General rights
Unless other specific re-use rights are stated the following general rights apply:
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.
• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Read more about Creative commons licenses: https://creativecommons.org/licenses/

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
The new London–Lund Corpus 2

Key methodological challenges and innovations

Põldvere, Nele  Lund University  <nele.poldvere@englund.lu.se>
Johansson, Victoria  Lund University  <victoria.johansson@ling.lu.se>
Paradis, Carita  Lund University  <carita.paradis@englund.lu.se>

Abstract
A few years ago at ICAME, we presented a work-in-progress paper on the very first stages of compiling the new London–Lund Corpus 2 (LLC–2). This year, we are in the fortunate position of being able to present the final corpus and to reflect on the whole compilation process. We do this by describing and critically examining three main methodological challenges that we encountered during the compilation of the corpus.

LLC–2 is a half-a-million-word corpus of contemporary spoken British English, collected 2014–2019. Its size and design are comparable to that of the world’s first machine-readable spoken corpus, the London–Lund Corpus (LLC–1) with data mainly from the 1960s. With LLC–2, we now not only have a new spoken corpus, but also a corpus that gives researchers the opportunity to make principled diachronic comparisons of speech over the past 50 years and to detect change in communicative behavior among speakers. Through a critical discussion of the methodological challenges of compiling LLC–2, we propose solutions to overcoming the challenges, often in innovative ways, as well as facilitate more informed uses of the corpus in the future.

The compilation of LLC–2 included a number of different stages: the design of the corpus, ethical and legal considerations, recruiting participants, data collection, transcription of the recordings, markup and annotation procedures, and finally making the corpus accessible to the research community. Each stage presented its own methodological challenges. The first challenge discussed in this paper concerns
the design of LLC–2. More specifically, the challenge was to strike a balance between LLC–2 as a representative collection of contemporary spoken English and its usefulness for diachronic comparisons with speech in LLC–1. According to Leech (2007), representativeness and comparability are inherently incompatible in corpus compilation since improvements in one lead to loss in the other. To achieve a sufficiently satisfactory balance between the two concepts, LLC–2 contains the same speech situations as LLC–1, i.e., primarily dialogue (in particular everyday face-to-face conversation) but also monologue; however, the specific recordings added to LLC–2 also reflect the technological advances of the past few decades, particularly with respect to speech situations such as telephone calls (e.g., Skype), and broadcast discussions and interviews (e.g., podcasts). As a result, LLC–1 and LLC–2 are comparable in the sense that they differ from each other in terms of only one parameter, the parameter of time, but, on its own, LLC–2 is also representative of the different kinds of communication channels used in the 21st century.

Another challenge concerned data collection. The distribution of key demographic categories such as age and gender in LLC–2 is relatively even. For example, there is only slight skewness in the data towards speakers aged between 35 and 59 years old. Achieving this balance was, however, not easy because of the large number of speakers needed to complete the corpus, the time constraints of the project, and the nature of the task at hand (i.e., to record a 30-minute conversation with other people). Therefore, our approach to data collection was largely opportunistic, meaning that we did not actively seek out recordings from certain groups of speakers, but instead accepted recordings from all speakers willing to be recorded for inclusion in the corpus. However, when it became clear at later stages of the data collection that some demographic categories were heavily skewed towards certain groups of speakers, efforts were made to reduce the bias to the extent possible. Moreover, the inclusion of online recordings in LLC–2 such as broadcast discussions and interviews allowed us to target specific demographic groups more easily, largely thanks to the wealth of data found online.

The third challenge discussed in this paper concerned transcriptions and markup. More specifically, it was important to develop a transcription and markup scheme that at the same time was (i) economical, (ii) useful for a wide range of areas in corpus
linguistics, and (iii) compatible with modern corpus linguistic and natural language processing tools. Therefore, the transcriptions in LLC–2 are orthographic and involve a manual transcription of words together with markups of basic spoken features. The transcription and markup scheme was kept as simple as possible, largely in consideration of the workload of the transcribers. Moreover, the scheme draws heavily on the guidelines of the International Corpus of English (ICE), which have proved useful for investigations of spoken English in a wide range of areas in corpus linguistics such as lexicology, morphosyntax and discourse analysis. In contrast to ICE, however, the transcriptions in LLC–2 are based on the standardised markup language XML, which is compatible with many of the well-known text processing tools currently used by corpus linguists such as AntConc and Wordsmith Tools. A feature of LLC–2 that is entirely new is that the transcriptions are time-aligned with the audio files, and they are released together in order to allow users to extend the orthographic transcriptions relative to their own research interests such as for prosodic research (see Põldvere et al., 2020 for the main challenges encountered in the public release of the LLC–2 audio material such as the anonymisation of the audio files).

In sum, the three main challenges and the innovations related to the compilation of LLC–2 concerned (i) the design of the corpus, (ii) data collection, and (iii) developing an effective and flexible transcription and markup scheme. As a critique of the solutions proposed in each stage, this paper serves as an example of how researchers may address recurring issues of spoken corpus development and analysis in the future.

References
Leech, Geoffrey. 2007. “New resources, or just better old ones? The Holy Grail of representativeness.” In Marianne Hundt, Nadja Nesselhauf and Carolin Biewer (Eds.), Corpus Linguistics and the Web (pp. 133–149). Rodopi.