RiCL Research in Corpus Linguistics

Introduction: Innovation in spoken corpus linguistics[1]

Robbie Love
Aston University / United Kingdom

Abstract – Over the decades, technological advancements have substantially improved the efficiency and scope of spoken corpus compilation, but there remain many challenges —both practical and theoretical— that constrain 1) the quality of spoken corpus data, 2) the scale to which spoken corpora can be compiled, and 3) the authenticity with which spoken language is represented in textual form. This special issue presents eight studies which address contemporary innovations in spoken corpus design, data collection, processing, and analysis, covering a range of speech contexts and varieties. The studies focus on registers including online workplace meetings, casual conversation, oral histories, oral proficiency interviews, and *YouTube* vlogs. Innovations include the integration of automated transcription tools, multimodal annotation schemes, creative participant recruitment methods, and developments in natural language processing (NLP). Three contributions offer critical reconceptualisations of traditional approaches to spoken corpus design, proposing strategies to improve the authenticity of spoken corpora.

Keywords – spoken corpora; corpus design; corpus construction; transcription; representativeness

Corpora derived from recordings of spoken language have long presented unique challenges from the perspectives of corpus design, compilation, processing, annotation, and analysis, among others. Early spoken corpora, such as the 500,000-word *London-Lund Corpus* (LLC; Greenbaum and Svartvik 1990), came about as the result of decades of labour-intensive, manual preparation of transcripts derived from analogue audio recordings. Since then, technological innovations have revolutionised the compilation of spoken corpora, and researchers have, over time, incrementally improved various stages of the corpus compilation pipeline to the benefit of the speed, scale, and diversity of spoken corpus compilation. Among the many innovations in this regard are the development of part-of-speech taggers trained on spoken data —e.g., the *British National Corpus 1994* (BNC1994; BNC Consortium 2007)— the creation of standard mark-up schemes for spoken texts —e.g., the *International Corpus of English*

[1] I am grateful to Carlos Prado-Alonso and Paula Rodríguez-Puente for their editorial advice and support, and to the 23 reviewers who provided double-blind anonymous peer review for the submissions to this special issue.

(ICE-GB; Nelson *et al.* 2002), the adoption of digital recording devices —e.g., the *British National Corpus 2014* (BNC2014; Love *et al.* 2017)— the use of crowdsourcing techniques for data collection —e.g., the *National Corpus of Contemporary Welsh* (CorCenCC; Knight *et al.* 2021)— and the time-alignment and anonymisation of audio files —e.g., the *London-Lund Corpus 2* (Poldvere *et al.* 2021).

Despite how far things have come, a number of challenges (both practical and theoretical) persist that constrain 1) the quality of spoken corpus data, (2) the scale to which spoken corpora can be compiled, and 3) the authenticity with which spoken language is represented in textual form. The papers in this special issue of *Research in Corpus Linguistics* represent some of these current challenges and the innovative solutions proposed to overcome them, which reflect, among other developments, the recent mass proliferation of artificial intelligence tools and the prominence of digitally-mediated spoken communication in day-to-day life. Collectively, the papers in this issue represent innovations in spoken corpus design (multimodal corpora, multilingual corpora, and data authenticity), construction (participant recruitment, automated transcription, and transcription of non-standard varieties), and analysis (comparability, sub-sampling, and manual coding schemes).

The first paper —by **Anne O'Keeffe, Dawn Knight, Geraldine Mark, Christopher Fitzgerald, Justin McNamara, Svenja Adolphs, Benjamin Cowan, Tania Fahey Palma, Fiona Farr, and Sandrine Peraldi**— introduces the *Interactional Variation Online* (IVO)[2] project and describes the compilation of a multimodal corpus of online workplace communication to facilitate analysis of verbal and non-verbal interactional features in virtual meetings. The project is timely in that it responds to a step-change in workplace practices in the wake of the COVID-19 pandemic, during which online meetings became much more common. The paper provides a replicable framework for multimodal corpus construction and describes the major stages in the design, collection, processing, and annotation of audiovisual data. Innovations include the unintrusive use of participants' own hardware to capture data, the integration of speech-to-text technology (*Otter.ai*)[3] to semi-automate the process of transcription, and the subsequent processing of the *Otter* transcripts using *ELAN* (Wittenburg *et al.* 2006). O'Keeffe *et al.* demonstrate how a user-driven model of

[2] https://ivohub.com/
[3] https://otter.ai/

corpus compilation, in which end-users are involved in the co-construction of the corpus design, can maximise the authenticity and utility of the resulting corpus.

The second paper discusses the design and compilation of another new spoken corpus. **Elizabeth Hanks, Tony McEnery, Jesse Egbert, Tove Larsson, Douglas Biber, Randi Reppen, Paul Baker, Vaclav Brezina, Gavin Brookes, Isobelle Clarke, and Raffaella Bottini** outline the development of the *Lancaster-Northern Arizona Corpus of Spoken American English* (LANA-CASE), a nationally representative corpus of spoken American English conversation (and American counterpart to the Spoken BNC2014). In this paper, the authors focus specifically on the earlier stages of corpus compilation, namely corpus design, participant recruitment, and data collection. In terms of corpus design, the authors draw upon Egbert *et al.* (2022) to describe the operational domain and develop an iterative sampling frame based on five selection criteria: age, race/ethnicity, gender, geographical region, and residential setting. The paper evaluates the effectiveness of a range of participant recruitment strategies, including innovative use of social media (e.g., *TikTok*), incentives for university students, and targeted outreach to specific populations such as older speakers and speakers from underrepresented racial/ethnic backgrounds. The data collection procedure makes innovative use of online survey platform *Phonic*[4] for the submission of vocal samples to aid speaker attribution. In conclusion, Hanks *et al.* emphasise the role of creative problem solving in addressing challenges in spoken corpus compilation and offer their solutions to these challenges as inspiration for future corpus compilers.

The third paper —by **Sarah Moeller, Alexis Davis, Wilermine Previlon, Michael Bottini, and Kevin Tang**— provides an example of the creation of a spoken corpus from existing audio-recorded data, namely oral histories. The paper describes the ongoing creation of a time-aligned, linguistically annotated corpus of *African American Language* (AAL) using oral histories from the *Joel Buchanan Archive of African American Oral History* (JBA).[5] When completed, the corpus is expected to comprise 500 oral histories interviews, representing AAL as spoken in southeast USA. This initiative aims to address the gap in accessible AAL data for linguistic research, which has implications for improving the performance of natural language processing technologies (NLP) —such as automatic speech recognition (ASR)— that are said to be

[4] https://www.phonic.ai/
[5] https://ufdc.ufl.edu/collections/ohfb

insufficiently trained on minority varieties. Moeller *et al.* discuss challenges associated with compiling a corpus from data not originally collected for the purposes of linguistic research, including a) the revision of pre-existing transcripts that were found to contain misrepresentation of AAL features not captured by standard orthographic conventions, and b) time-alignment of the audio recordings and corresponding transcripts, using the toolkits *Aeneas* (Pettarin 2017) and the *Montreal Forced Aligner* (MFA; McAuliffe *et al.* 2017). A case study, based on a small sub-set of transcripts, demonstrates efforts to create tools that can automatically tag and align AAL features (e.g., habitual *be*, multiple negation), with the ultimate goal of improving NLP systems for AAL while also preserving the rich cultural narratives found in African American oral histories.

In the fourth paper, **Nicholas Smith, Cristiano Broccias, and Cathleen Waters** offer a critical evaluation of the comparability of the two iterations of the *Spoken British National Corpus* (BNC) from the 1990s and 2010s. Focussing on the past perfect (e.g., *That's the first time you'd met her?*), the authors evaluate the suitability of these corpora for studying sociolinguistic variation and change over time. The paper identifies key issues such as differences in transcription quality, annotation standards, and sampling methods between the two corpora. To address these issues, Smith *et al.* propose modifications to the *BNClab* subcorpus (Brezina *et al.* 2018), which balances the demographic variables gender, age, socio-economic status and region across the two periods. The modified sample (*BNClab-M*) reduces the number of demographic variables and speakers in order to boost comparability. A case study on the past perfect and its variants, including non-standard forms, finds that while there has been a significant increase in the use of the past perfect in recent British English conversation (contradicting the findings of Bowie *et al.* 2013 and Smith and Waters 2019), sociolinguistic patterns of variation remain complex. The study offers methodological insights for improving a) the quality of corpus comparability, and b) the precision and recall of grammatical constructions, and provides implications for both corpus researchers and language teachers.

The fifth paper offers another critical evaluation of spoken corpus design, this time in the context of learner corpora. **Pascual Pérez-Paredes and Geraldine Mark** critically examine the use of interviews in the compilation of spoken learner corpora, drawing distinction between conceptualisations of the interview as both an elicitation technique on the one hand, and a distinct genre on the other. They argue that, despite

often being used as a benchmark for spoken learner language, interview data (especially that derived from oral proficiency assessments) may not provide an authentic representation of everyday spoken learner language. In a series of case studies on the use of adverbs across speakers from four first language (L1) backgrounds in the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin *et al.* 2010) and the *Louvain Corpus of Native English Conversation* (LOCNEC; De Cock 2004), the paper explores the role of interviewers in influencing the quality and nature of learner data and suggests that interviews often lack interactional features of natural conversation, such as co-construction, turn-taking, and back-channelling. The paper calls for a reconceptualisation of how interviews are used in learner corpus research, recommending that future research designs consider alternative methods for gathering authentic spoken learner data. Pérez-Paredes and Mark advocate for a more critical reflection on the comparability and representativeness of learner corpora, especially in terms of interactional features that are characteristic of everyday spoken language.

Continuing the theme of spoken learner corpora, the sixth paper —by **Yejin Jung, Dana Gablasova, Vaclav Brezina, and Hanna Schmück**— presents a novel coding scheme designed to identify and classify linguistic expressions of opinion in second language (L2) interactive spoken English. The research addresses a gap in existing annotation frameworks, which tend to focus on written language or first language use. The paper discusses challenges in recognising and quantifying evaluative language, particularly in spoken interaction, whereby opinions are often co-constructed between speakers. The coding scheme proposed in the study is applicable in language teaching and assessment contexts, allowing researchers to measure the frequency and complexity of opinion statements, while recording L2 learners' ability to state and support opinions independently. The scheme categorises opinion statements into simple and complex forms, the latter including supporting statements such as reasons, elaborations, or evidence. The study evaluates the reliability of the coding scheme on a sample of 29 texts from the *Trinity Lancaster Corpus* (TLC; Gablasova *et al.* 2019), which contains transcripts from Trinity College London's Graded Examinations in Spoken English (GESE). Jung *et al.* demonstrate that the scheme offers a resource for investigating evaluative language as a component of the pragmatic abilities of L2 learners.

In another critical reflection, the seventh paper —by **Giorgia Troiani, John W. Du Bois, and Andrey Filchenko**— advocates for an alternative approach to spoken

corpus design, in which priority is given to representation of participants' lives as opposed solely to the representation of spoken output. The authors critique the reliance on discourse spontaneity as a criterion for corpus design, arguing that 'spontaneous' data may still display artificial interactional dynamics. Through the lens of the 'cast the net wide' framework, first implemented in design of the *Santa Barbara Corpus of Spoken American English* (SBCSAE; Du Bois *et al.* 2000) and adapted for the *Multimedia Corpus of Modern Spoken Kazakh Language* (MULTICORSKL; Filchenko *et al.* 2023), the paper distinguishes between 'spontaneous' and 'naturally occurring' discourse, arguing that the latter —language used in speech events that are socially and interactionally relevant for the participants, and not imposed by researchers— offers a more faithful reflection of the speakers' real lives. Drawing on examples from Kazakh, Italian, Bustocco, Mixtec, and English, the study explores the consequences of the data collection process, showing how interactional features like backchanneling and turn-taking vary according to the nature of the event and the research protocols. The authors propose innovative adjustments to corpus design to focus on participant agency and the integration of naturally occurring events to facilitate the development of corpora that reflect both language and lived experiences.

In the eighth and final paper, **Hülya Mısır** describes the design and construction of a multilingual corpus of Turkish social media influencers' *YouTube* vlogs. The paper discusses the challenges of transcribing and annotating vlog content. An evaluation of the suitability of *YouTube's* auto-generated captions as the basis for corpus transcripts found that, in the case of Turkish, the quality of *YouTube's* ASR was insufficient to offer better efficiency than manual transcription, so the latter was used. Mısır then describes the use of *ELAN* (Wittenburg *et al.* 2006) to develop a bespoke annotation system for the influencers' translanguaging practices, which facilitates representation of a range of translanguaging categories, including the integration of foreign language items, digital lexis, and multimodal resources such as emojis and visual elements. The corpus contains over 120,000 tokens of transcribed speech, offering a resource for examination of the translanguaging practices and multimodal communication of influencers. The paper concludes by describing the ethical principles applied in the collection of data from *YouTube* and arguing that there is a need for traditional transcription conventions to evolve to adapt to multimodal digital communication, especially in the context of translanguaging.

The papers in this special issue are indicative of just some of the current trends in (spoken) corpus linguistics, which seeks to become more multimodal, more linguistically diverse, and more authentic. As technology has advanced, so too have the methods and tools for compiling and analysing spoken corpora, which capture increasingly diverse contexts, registers, and language varieties. It is my hope that the papers in this issue will provide inspiration for the next generation of innovations in spoken corpus linguistics.

REFERENCES

BNC Consortium. 2007. *The British National Corpus, XML Edition*. Oxford Text Archive: http://hdl.handle.net/20.500.14106/2554.

Bowie, Jill, Sean Wallis and Sebastian Aarts. 2013. The perfect in spoken British English. In Sebastian Aarts, Joanne Close, Geoffrey Leech and Sean Wallis eds. *The Verb Phrase in English: Investigating Recent Language Change with Corpora.* Cambridge: Cambridge University Press, 318–352.

Brezina, Vaclav, Dana Gablasova and Susan Reichelt. 2018. *BNClab*. http://corpora.lancs.ac.uk/bnclab

De Cock, Sylvie. 2004. Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures* 2: 225–246.

Du Bois, John W., Wallace L. Chafe, Charles Meyer, Sandra A. Thompson and Nii Martey. 2000. *Santa Barbara Corpus of Spoken American English*. Philadelphia: Linguistic Data Consortium.

Egbert, Jesse, Douglas Biber and Bethany Gray. 2022. *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*. Cambridge: Cambridge University Press.

Filchenko Andrey, Giorgia Troiani, John W. Du Bois, Gulnar Sarseke, Akyl Akanov, Moldir Bizhanova, Nikolay Mikhailov, Tansulu Temirbekova, Bybaris Seitak and Zhansaya Turaliyeva. 2023. *Multimedia Corpus of Spoken Kazakh Language* (version 1).

Gablasova, Dana, Vaclav Brezina and Tony McEnery. 2019. The Trinity Lancaster Corpus: Development, description, and application. *International Journal of Learner Corpus Research* 5/2: 126–158.

Gilquin, Gaëtanelle, Sylvie De Cock and Sylviane Granger. 2010. *The Louvain International Database of Spoken English Interlanguage. Handbook and CD-ROM*. Louvain-La-Neuve: Presses universitaires de Louvain.

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.

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

Knight, Dawn, Fernando Loizides, Steven Neale, Laurence Anthony and Irena Spasić. 2021. Developing computational infrastructure for the CorCenCC corpus: The

National Corpus of Contemporary Welsh. *Language Resources & Evaluation* 55: 789–816.

McAuliffe, Michael, Michaela Socolof, Michael Wagner and Morgran Sonderegger. 2017. Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *INTERSPEECH:* 498–502.

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.

Pettarin, Alberto. 2017. *Aeneas: Automagically Synchronize Audio and Text*. https://www.readbeyond.it/aeneas/

Põldvere, Nele, Victoria Johansson and Carita Paradis. 2021. On The London–Lund Corpus 2: Design, challenges and innovations. *English Language and Linguistics* 25/3: 459–483.

Smith, Nicholas and Cathleen Waters. 2019. Variation and change in a specialized register: A comparison of random and sociolinguistic sampling outcomes in Desert Island Discs. *International Journal of Corpus Linguistics* 24/2: 169–201.

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, Josheph Mariani, Jan Odijk and Daniel Tapias eds. *Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genoa: European Language Resources Association, 1556–1559.

Corresponding author
Robbie Love
Aston University
School of Law and Social Sciences
Birmingham
B4 7ET
United Kingdom
Email: r.love@aston.ac.uk